AD________________

GRANT NUMBER DAMD17-93-J-3007

TITLE: Implementation of Computer Assisted Breast Cancer Diagnosis

PRINCIPAL INVESTIGATOR: Shih-Chung Lo, Ph.D.

CONTRACTING ORGANIZATION: Georgetown University
Washington, DC 20057

REPORT DATE: July 1996

TYPE OF REPORT: Final

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE July 1996	3. REPORT TYPE AND DATES COVERED Final (1 Dec 92 - 30 Jun 96)

4. TITLE AND SUBTITLE Implementation of Computer Assisted Breast Cancer Diagnosis

5. FUNDING NUMBERS
DAMD17-93-J-3007

6. AUTHOR(S)
Shih-Chung Lo, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Georgetown University
Washington, DC 20057

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick
Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

This project aims at the implementation of a computer-aided diagnosis system for the detection of microcalcifications on mammograms based on the algorithms developed by the principal investigator and others. In addition, the proposed research includes: (1) algorithm improvement for the detection of microcalcifications, (2) mammographic image compression and its impact on computer-aided diagnosis (CADx), and (3) computer-aided classification of benign and malignant masses on mammograms.

In the past three years, we have developed several algorithms and have studied part of the proposed research: (a) development of filtering techniques with wavelet transform to reduce mammographic structures other than microcalcifications, (b) performance of preliminary study in the detection of microcalcifications, (c) performance of mammographic compression studies using split gray values in conjunction with wavelet and full-frame discrete cosine transform (DCT) techniques, (d) evaluation of the impact of the compression with respect to various degrees of data compression, and (f) implementation of CADx system in a DEC Alpha workstation.

14. SUBJECT TERMS Breast Cancer, Diagnosis, Computer

15. NUMBER OF PAGES
146

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

7/29/96

PI - Signature Date

The Final Report for Project Titled:

Implementation of Computer Assisted Breast Cancer Diagnosis

(US Army Grant No. DAMD17-93-J-3007)

TABLE OF CONTENTS

The Final Report for Project Entitled:

Implementation of Computer Assisted Breast Cancer Diagnosis

(US Army Grant No. DAMD17-93-J-3007)

Army grant DAMD17-93-J-3007 was initiated in December 1992 for completion in December 1995. This was extended to a completion date of June 1996. This represents the final report for this project.

1. Introduction

Recently, several investigators have proposed a number of methods for the automatic detection of microcalcifications and masses on mammograms. Significant improvements in accuracy have been made since the initial attempt [Chan 1987; 1988] to apply the computer algorithms for the detection of microcalcifications. We believe that it is important to implement the program into a high speed workstation and conduct a large scale clinical trial in order to evaluate its clinical practicability and limitations. Although the false-positive rate for the detection of masses is still very high, we have been using an artificial neural network to classify malignant and benign masses. We believe that the creation of a computer program to analyze features of suspected masses will give rise to a more useful and fundamental approach to computer-aided diagnosis.

Because digital mammography produces a large data volume for its high-resolution imaging, data compression is an important means to facilitate the mammographic image transmission and storage. We have studied characteristics of the mammograms and developed compression methods specifically for mammograms using gray value splitting in conjunction with wavelet and full-frame discrete cosine transform (DCT) techniques. Effects of applying the data compression to the proposed computer aided diagnosis (CADx) scheme in the detection of microcalcifications were also tested during this reporting period.

2. Research in the Detection of Microcalcifications

2.1. Detection of Suspected Microcalcifications

Microcalcifications in breast cancer are reported to occur with five or more microcalcifications as a cluster in a 1cm^2 area [Black 1965, Fisher 1975]. When the digitization pixel size is 50 µm (using a Lumiscan 150), there are 40,000 pixels in a 1cm^2 area. To have five detections or pixels (0.0125%) possessing high intensity in the area means that one should set a threshold on pixel intensity of approximately 3.61 σ (σ: standard deviation). In one experiment, we used 3.02 σ as the threshold corresponding to a maximum of 50 pixels (0.125% as indicated in Figure 1) due to a potentially larger microcalcification containing several detected pixels together. Note that a background trend correction was applied to each image block prior to the statistical calculation. The previously detected suspected areas (i.e., 50 pixels) were masked with the mean value in this detecting procedure. This procedure was performed with a $1 \mathrm{cm}^2$ template (200×200 pixels) by moving 190 pixels per step for each operation and by scanning through the mammogram horizontally and then vertically.

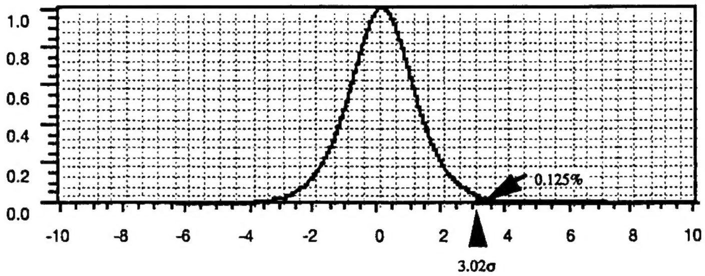


Figure 1. Assuming the noise spectrum fits Gaussian distribution, only 0.125% of pixels have an intensity higher than 3.02 σ.

After carefully evaluating twenty-two mammograms containing subtle microcalcifications (only three clustered microcalcifications on three mammograms were associated with malignant process), we found that the use of 3.02 σ for the threshold value was fine except for radiolucent regions (OD > 2.3) where a threshold value should be set at 2.75 σ corresponding to 120 pixels (0.3%) in a $1 \mathrm{~cm}^2$ area. In addition, when a large area was detected (> 30 pixels) then additional pixels corresponding to the area would be granted in the local operation. Our results indicated that all microcalcifications (27 clusters confirmed by biopsy and 126 singles were confirmed by an experienced radiologist) were detected through the above procedure. However, an average of 858 suspected areas per mammogram was obtained (i.e., 99.5% false-positive rate for 100% true-positive detection). This procedure is equivalent to a pre-scan process of a computer-aided diagnosis in the detection of microcalcifications [Chan 1987; 1990]. The important point here is that we have developed an effective computer program that can detect all microcalcifications. It takes 5-7 seconds on a DEC Alpha computer to run a digital mammogram of $4{,}096 \times 5{,}120$ pixels. The suspected areas will be used for the further evaluation of CADx using more stringent criteria and in the mammographic image compression for error handling in the next section.

3. Adaptive Lossless Mammographic Image Compression

We have also developed an adaptive lossless compression scheme for mammograms by combining a high compression method and techniques involving the detection of all suspected microcalcifications to ensure data accuracy in the clinically significant areas. In the previous section, we described how to detect suspected microcalcifications. It is no a big task to handle 858 suspected areas when compared to the compression of a 4K×5K mammogram. However, we can preserve the maximum data accuracy on clinically significant areas. This type of error control should be used in any medical image compression scheme when possible.

3.1. Mammographic Image Compression via Wavelet Decomposition

Recently, we have used a wavelet transform for mammographic image compression [Daubechies 1988, Mallat 1989, Cody 1992, Atonini 1992]. Before the wavelet transform, the boundary of the breast was outlined. Only the area within the boundary was the area to be compressed. Figure 2 shows a typical multi-level wavelet transform and the associated compression procedure. The larger the image, the more levels of wavelet transform can be applied. In general, "A" contains a much smaller computer space than "B" and "A" space + "B" space is about 4K×5K×3 bit (a compression ratio of 4:1). If the air region is included in the compression process, the average error-free compression ratio is ≈2.5:1.

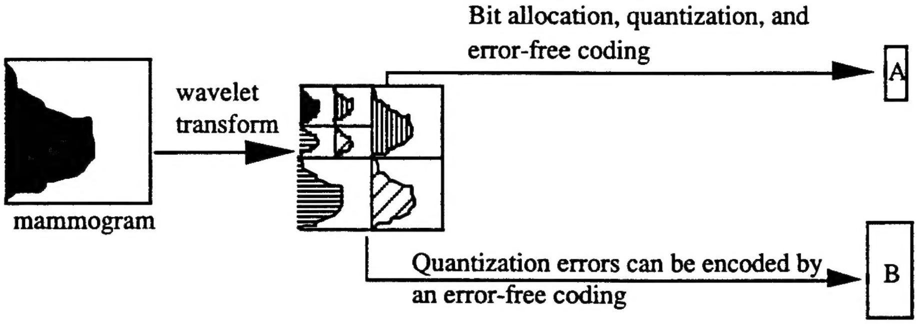

Bit allocation, quantization, and error-free coding
A
wavelet transform
mammogram
Quantization errors can be encoded by an error-free coding
B

Figure 2. A typical wavelet decomposition and associated compression procedure for a mammogram. (Note: only a two-level decomposition is shown.)

In this study, we decomposed each image with 7-level wavelet transform; hence, the smallest size image will be a matrix of 128×160 pixels. The lowest resolution subimage will be further decomposed by an operation called deferential pulse code modulation (DPCM). The entropy of the all-decomposed subimages will be calculated to determine the best wavelet kernel for the mammographic image compression.

3.2. Error-Controlled Compression for Digital Mammograms

We believe that an accurate error-control procedure is an innovative solution to make a compression scheme clinically useful. A computer scheme for the compression was tested and is described as follows:

(a) Detect all suspected microcalcifications (clusters and singles) based on the method described in Section 2.

(b) Perform an error-free compression using DPCM and arithmetic coding on the detected areas. Replace the area with surrounding intensity using cubic spline interpolation.

(c) Perform multi-level wavelet transform for the mammogram.

(d) Perform quantization on the wavelet domain (For the higher level of low resolution subimages the less destructive quantization should be applied.)

(e) Perform an entropy coding on quantized subimages to get file "A" indicated in Figure 2. (arithmetic coding [Witten 1987] for uncorrelated coefficients and L-Z coding [Ziv 1978] for correlated data sequence).

3.3. Experimental Results

The unique point of this work is to add the error-free feature for the suspected disease areas to a compression scheme. No compression artifact shall be observed by an experienced breast radiologist. One must realize that there is no need to digitize a resolution as high as 50μm/pixel except those areas containing subtle microcalcifications. However, the error control feature reduced some degrees of the entire compression efficiency (ratio). Equation (1) provides a formula to calculate the effective compression ratio when the error-control feature is added into the compression system:

$$R_t = \frac{R \times R_e \times T}{(R - R_e) \times N \times S + R_e T} \qquad ...(1)$$

where T is the total number of pixels in the original mammogram, S is the number of pixels in the suspected area for error-free coding, N denotes the number of suspected areas, R is the compression ratio obtained by performing a transform (wavelet) coding, R_e is the average compression ratio to encode microcalcification areas losslessly, and R_t is the total effective compression ratio.

We tested the same twenty-two mammograms as used in Section 3. We calculated the effective compression ratio by providing values:

$N \approx 858$;

$S \approx 640$ ($\approx 25 \times 25$ pixels) which was averaged from 81% tiny suspects requiring 20×20 pixels (i.e., $1\text{mm} \times 1\text{mm}$ area) and 19% medium-sized suspects requiring 40×40 pixels;

$T = 20{,}971{,}520$ ($4{,}096 \times 5{,}120$);

$R_e \approx 2.5$;

$R \approx 40{:}1$ (estimated acceptable compression ratio) which is partly due to the fact that $\approx$50% of mammogram contains air space.

Substituting the above values into Equation (1), we received $R_t \approx 29$ which also indicates that an additional 40% of the compressed data was increased when the error-free feature was added to the compression scheme. Since each 12-bit datum is stored in a 16-bit computer space, R_t was 38 for current commercial data systems. Because the suspected areas may contain significant clinical information, we believe that the error control feature is necessary and is a cost-effective approach for mammography data reduction.

4. Recognition of Mammographic Microcalcifications with an Artificial Neural Network

4.1. Detection of Clustered Microcalcifications

We have developed a computer-aided diagnosis (CADx) program for automated detection of clustered microcalcifications in digital mammograms. In this study, we investigated the use of a convolution neural network (CNN) in conjunction with the CADx program to reduce false-positive (FP) detections.

Screen-film mammograms containing subtle microcalcifications were digitized with a laser film scanner. After signal-to-noise ratio (SNR) enhancement and background removal with a spatial filter, potential signal sites were detected with a locally adaptive gray-level thresholding technique. The size and contrast were used to discriminate false signals from true microcalcifications. The remaining signals were then inspected by the CNN. Image blocks containing individual microcalcifications in the SNR-enhanced images were input to the CNN. The CNN consisted of nodes organized in groups and the weights connecting the nodes were organized by convolution kernels. These weights integrated neighborhood information for recognition of the true signals. After training, we found that a CNN with two hidden layers, both containing 10 groups of nodes, was effective in the classification of true and false signals. The output signals from the CNN further underwent a regional clustering algorithm for detection of clustered microcalcifications.

We found that the CNN could classify individual microcalcifications with the area under the ROC curve, Az, of 0.88. Free-response ROC (i.e., FROC) analysis showed that the addition of CNN classification to the CADx program reduced the false-positive cluster detection by 60-70% for a given true-positive (TP) rate. After adding a criterion regarding a minimum of three calcifications in one cluster for a detection, the Az was increased to 0.96. These results indicate that the CNN can significantly increase the accuracy of the CADx program.

4.2. Classification of Malignant and Benign Clustered Microcalcifications

We have developed computer vision methods for classification of malignant and benign clustered microcalcifications. Mammograms are digitized at a pixel size of 35 mm × 35 mm. The program operates locally in regions of interest (ROIs) containing clusters of microcalcifications on the mammograms. Morphological features characterizing the microcalcifications and texture features characterizing textural changes in the tissue region surrounding the cluster are extracted from the ROIs. For extraction of texture features, we first employ a distance-weighted interpolation technique to estimate the low-frequency background of the ROI using a band of pixels around its perimeter. The spatial gray level dependence (SGLD) matrices of the background-corrected ROI are determined at various pixel pair distances. Thirteen texture features that characterize the ROI, such as correlation, energy, inertia, inverse difference moment, and entropy, are calculated from the SGLD matrices.

For extraction of morphological features, a segmentation method is used similar to that in the detection program except that segmentation is applied to an unfiltered image in order to avoid distortion of its shape due to signal-to-noise ratio (SNR) enhancement. An ROI containing a microcalcification is background-corrected and the signal is extracted based on the local SNR using a region growing technique. We calculate visibility descriptors such as the SNR, mean density, and size of the microcalcifications, shape descriptors such as the second moments, the ratio of the second moments, the eccentricity and the ratio of major and minor axes of an effective ellipse, and determine cluster features such as the standard deviation, the maximum, and the coefficient of variations of the visibility descriptors, shape descriptors, and the number of microcalcifications within the cluster. We have trained a linear discriminant classifier (LDA) to classify the input features. The performance of the trained classifier has been tested both with a jackknife method and a cross-validation method. Both methods yielded similar test results. The discriminant scores of the LDA were analyzed with Receiver Operating Characteristic (ROC) methodology and the area under the ROC curve (Az) was used as a performance index. In the texture feature space, the LDA classifier achieved an Az of 0.88 for training and 0.84 for testing. In the morphological feature space, the LDA classifier achieved an Az of 0.84 for training and 0.79 for testing. In the combined texture and morphological features, the Azs were improved to 0.94 and 0.89, respectively, for training and testing. We have also trained a non-linear classifier, a back-propagation neural network (BPN), to classify the malignant and benign microcalcifications. In the texture feature space, the BPN classifier achieved an Az of 0.88 for training and 0.86 for testing. In the morphological feature space, the BPN classifier achieved an Az of 0.84 for training and 0.80 for testing. In the combined texture and morphological features, the Az's were improved to 0.94 and 0.91, respectively, for training and testing. These results demonstrate the feasibility of our approach to classification of malignant and benign microcalcifications.

5. Recognition of Mammographic Masses

5.1. Detection of Mammographic Masses

(A) Computerized detection of masses on mammograms

We have developed a new approach for segmentation of suspicious mass regions on digitized mammograms using an adaptive Density-Weighted Contrast Enhancement (DWCE) filter in conjunction with Laplacian-Gaussian (LG) edge detection. The DWCE filter can enhance masses of a wide range of intensities and sizes, and suppress background intensity variations. The algorithm processes a mammogram in two stages. In the first stage the entire mammogram is filtered globally using a DWCE adaptive filter which enhances the local contrast of the image based on its local mean pixel values. The enhanced image is then segmented with an LG edge detector into isolated objects. A feature classifier using morphological or texture features is used to reduce the number of FPs. In the second stage of processing, the DWCE adaptive filter and the edge detector are applied locally to each of the segmented object regions detected in the first stage. The local operation allows more precise extraction of the features of the objects. The number of objects is further reduced based on these features. ROIs are extracted from the image based on the remaining object set. The selected ROIs are input to either an LDA classifier or a convolution neural network to further differentiate TPs and FPs as described below. Using a cross-validation test method with two partitions, our results indicated that the current algorithm achieved an average test TP rate of 80% at about 2.1 FPs/image and a TP rate of 90% at 4.7 FPs/image. This accuracy may not be adequate in clinical practice, however, it demonstrates the feasibility of detecting masses on mammograms with the new DWCE technique. We therefore propose to evaluate the performance of the algorithm in a preclinical trial using a large number of randomly selected clinical cases. The causes of FP detections in such a test will be analyzed, and more effective FP reduction methods will be developed in order to improve the detection accuracy.

(B) Multiresolution wavelet decomposition and texture analysis

We have developed a new method to distinguish abnormal from normal tissue for CAD algorithms using texture analysis. An ROI containing mass or normal breast tissue is input to the program. The wavelet transform is used to decompose the ROI into several scales. Global multiresolution texture features are calculated from the SGLD matrices of the low-pass wavelet coefficients up to a certain scale and then at variable distances between the pixel pairs. Texture features in the suspicious object sub-region and their differences with features in the peripheral sub-regions of the ROI are also calculated to form a local texture feature space. Stepwise linear discriminant analysis is used to select effective features from the combined global-local feature space to maximize the separation of mass and normal tissue ROIs. To evaluate the accuracy of this method, we used 168 ROIs containing a biopsy-proven mass and 508 ROIs with normal dense, mixed dense/fatty, or fatty tissues extracted from digitized mammograms by radiologists. The ROIs were randomly and equally divided into a training and a test group. It was found that, using the global multiresolution feature space alone, the Az was 0.89 and 0.87 for the training and test groups, respectively. Using local features only, the Az was 0.88 and 0.85 for the training and test groups, respectively. With the combined global and local feature spaces, the Az reached 0.95 and 0.91 for the training and test groups, respectively. When this classification method was applied to the false-positives detected by the automated mass detection program using the DWCE approach described above, the classification accuracy in terms of Az reached 0.97 during training and 0.96 during testing in the combined global and local feature space. The results demonstrate that an LDA using a combination of the global and the local texture features can effectively classify masses from normal tissue on mammograms. This classifier will be incorporated into the automated mass detection program as one of the steps to reduce FP detections in the preclinical trial.

(C) Artificial neural network

We have investigated the use of a convolution neural network (CNN) and a backpropagation neural network (BPN) for classification of ROIs on mammograms as either masses or normal tissue. A CNN is a BPN with two-dimensional weight kernels that operate on images. A generalized, fast and stable implementation of the CNN has been developed. ROIs containing masses and normal breast tissue are first segmented with an automated detection program. The CNN input images are obtained from the ROIs using (i) averaging and subsampling, and (ii) texture feature extraction from SGLD matrices and gray level difference statistics (GLDS) vectors on smaller sub-regions inside the ROI. In (ii), features computed over different sub-regions were arranged as texture-images, and subsequently used as inputs to the CNN. Input features to the BPN are obtained from SGLD matrices at multiple resolutions. Using 168 ROIs containing masses and 504 ROIs containing normal tissue, we found that the test Az reached 0.83 for the CNN using spatial input images, 0.87 using spatial and texture images, 0.88 for the BPN using SGLD texture features, and 0.91 for a combination of the CNN and BPN outputs. Our results indicate that the CNN performance may be improved by using additional texture information and that the overall performance may be improved by combining CNN and BPN classifiers.

(D) Feature selection

The performance of a feature classifier in a CAD scheme depends strongly on feature selection. For the LDA, we use a stepwise LDA procedure to select significant feature variables for the classification tasks. In order to have a general feature selection method that can be applied to both linear and non-linear classifiers, we have investigated the application of a genetic algorithm (GA) for feature selection. One of our applications is to select features for the classification of masses and normal breast tissue. ROIs containing biopsy-proven masses and normal ROIs containing breast parenchyma are first segmented from mammograms. A total of 587 texture and morphological features are automatically extracted from each ROI. Multiple regression is applied to the features selected by the GA to form a discriminant function with the training set. The presence/absence of a feature in the regression is coded by a 1 or 0 at the appropriate gene in a chromosome in the GA. The fitness and survival rate of a chromosome are determined by Az. The chromosomes are allowed to crossover, mutate, and evolve for a number of generations in a training procedure. The final selected features are used for classification of the test set. To evaluate the effectiveness of this GA, we used 168 ROIs with masses and 504 ROIs with normal tissue as our data set. We randomly divided the data set into 10 partitions of training and test subsets. The GA selected an average of 20 features from the 587 input features for each training process. It was found that the average training and test Az values reached 0.93 and 0.89, respectively. This accuracy is superior to that obtained with the entire feature set input to the classifier without feature selection, or that with features selected individually based on their distributions. We also compared the results to feature selection using the stepwise LDA method. Using the same cross-validation test technique, the test Az's obtained with both methods were similar, indicating that the GA and stepwise LDA approaches can provide near-optimal feature selection for linear classifiers.

5.2 Classification Of Malignant And Benign Mammographic Masses

We have investigated the classification of malignant and benign masses on mammograms. After ROIs containing suspicious masses are located by the automated mass detection program on the mammogram, segmentation and feature extraction are performed locally in each ROI. A new segmentation method has been developed by the research team based on a migrating mean clustering algorithm. An ROI is first corrected for the low-frequency structured background. This method then separates the mass from the surrounding background based on clustering of pixels with similar gray level and edge gradient information. The two groups of pixels are coded as a binary image so that a simple edge tracking algorithm can define the boundary. We extract morphological features such as the fuzziness or spiculation of the mass margin which is quantified by the root-mean-square (RMS) variation around a smoothed version of the edge, the perimeter-to-area ratio, and shape features such as circularity, rectangularity, ratio of its axes, and shape features derived from the normalized radial length. We also extract texture features in a 40-pixel-wide boundary region surrounding the mass from the SGLD matrices. The features are then input to an LDA or a BPN classifier to distinguish the malignant and benign masses. The results indicated that the migrating mean clustering method could extract mass margins more closely than other edge detection techniques that we tested. With the morphological features and texture features derived from the boundary regions surrounding the mass, we obtained a training Az of 0.86 and a test Az of 0.82 for a group of 85 malignant and 83 benign masses. The Az was 0.86 by a radiologist's visual evaluation in the same set of mammograms. This result is encouraging although improved methods still need to be developed to further increase the classification accuracy before clinical implementation.

6. Status Report in the Implementation of CADx for the Detection of Clustered Microcalcifications

We continue to work on the CADx program with a DEC Alpha workstation. The basic user interface is complete. The user interface can select a mammogram and display it on the workstation. Several image functions have been implemented: (1) "window and level" for the adjustment of the brightness and contrast, (2) pan, (3) a cursor box for the user to select the area of interest, (4) print the image with CADx marks on high quality paper or a laser film. Initial clinical trial began January 15, 1996 at the Breast Imaging Division of Georgetown University Hospital. The results of this study will be presented at the 1997 SPIE Medical Imaging Conference at Newport Beach, California.

7. Contractual (SOW) Issues

Dr. R.V. Shah, chief breast radiologist at Brooke Army Medical Center and Dr. Don Smith, attendant breast radiologist at Madigan Army Medical Center have sent us some proven cases (in the Spring of 1995) associated with mammographic microcalcifications for inclusion in our test database [Private Communication]. We are in the process of installing our software for evaluation at Army Hospitals. The CADx clinical trial can be started anytime when they are ready for the experiment. At present, radiologists would like to have an integrated viewing system so that thay can evaluate the effects of CADx in a clinical setting. We are currently negotiating with R2 Technology Inc. who have a product that synchronizes soft copy images on small monitors mounted under the bench of the mammography viewing alternator using a bar code system. We plan to miniturize the mammograms and provide marks indicated by our CADx. The images will be interfaced to the R2 monitors to facilitate the clinical use of this development. Dr. R.V. Shah can be reached at (210)916-4062. R2 Technology's phone number is (415)254-8988.

8. Conclusions

During the past three years, we have spent our effort not only in algorithm improvement but also in merging our newly developed algorithm in C and useful codes previously developed by Dr. Chan and her colleagues.

At this point, we have completed our mammographical image compression and CADx research in terms of algorithm improvement and computer speed. Database collection is underway and will continue in the clinical tests to be conducted. Several basic functions and user interface have been implemented in the workstation. The CADx clinical trial has been undertaken at Georgetown University Medical Center. We will report the results of the clinical test in a future paper.

References

Antonini M, Barlaud M, Mathieu P, Daubechies I: "Image Coding Using Wavelet Transform," IEEE Trans. Image Proc., vol. 1 No. 2, 1992, pp. 205 - 220.

Black JW, Young B: "A Radiological and Pathological Study of the Incidence of Calcifications in Diseases of the Breast and Neoplasms of Other Tissues," Br J Radiol 1965;38:596.

Chan HP, Doi K, Galhotra S, Vybomy CJ, MacMahon H, Jokich PM: Image Feature Analysis and Computer-aided Diagnosis in Digital Radiography. 1. Automated Detection of Microcalcifications in Mammography," Med. Phys., 1987;14:538.

Chan HP, Doi K., Vybomy CJ, et al.: "Improvement in Radiologists' Detection of Clustered Microcalcifications on Mammograms: The Potential of Computer-Aided Diagnosis," Invest. Radio. vol. 25, 1990, pp. 1102-1110.

Cody MA, "The Fast Wavelet Transform," Dr. Dobb's Journal, April 1992, pp. 16-28.

Daubechies I, "Orthonormal Based of Compactly Supported Wavelets", Comm. on Pure and Appl. Math., Vol. XLI, 1988, pp. 909-996.

Fisher ER, Gregorio RM, Fisher B, Redmond C, Vellios F, Sommers SC: "The Pathology of Invasive Breast Cancer," Cancer 1975;36:1.

MacMahon H, Doi K, Sanada S, Montner SM, Giger ML, Metz CE, Nakamori N, Yin F, Xu X, Yonekawa H, and Takeuchi H: "Data Compression: Effect of Data Compression on Diagnostic Accuracy in Digital Chest Radiography", Radiology, Vol. 178, No. 1, Jan. 1991, pp. 175-179.

Mallat S, "A Theory For Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Trans. Pat. Anal. Mach. Intel., Vol. 11 No. 7, 1989, pp. 674-693.

Swets JA and Pickett RM, Evaluation of Diagnostic Systems, Academic Press, New York, 1982.

Witten IH, Neal RM, and Cleary JG: "Arithmetic Coding for Data Compression," Comm. of the ACM, Vol. 30, June 1987, pp. 520-540.

Ziv J and Lempel A: "A Universal Algorithm for Sequential Data Compression," IEEE Trans. on Info. Theory, Vol. IT-23, No. 3, May 1977, pp. 337-343.

Presentations and Publications During the 2nd Year of the Project

1. Petrosian A, Chan HP, Helvie MA, Goodsitt MM, Adler DD: "Computer-aided diagnosis in mammography: classification of masses and normal tissue by texture analysis," Physics in Medicine and Biology 1994; 39: 2273-2288.

2. Cheng SNC, Chan HP, Helvie MA, Goodsitt MM, Adler DD, St. Clair D: "Classification of mass and non-mass regions on mammograms using artificial neural network," J. of IS&T 1994; 38: 598-603.

3. Lo SC, Kim MB, Li H, Krasner BH, Freedman MT, and Mun SK, “Radiological Image Compression: Image Characteristics and Clinical Consideration,” SPIE Proceedings, Medical Imaging 1994, vol. 2164, pp. 276-281.

4. Wu YC, Lo SC, Freedman MT, Zuurbier RA, Hasegawa A, Mun SK: “Classification Of Microcalcifications In Radiographs Of Pathological Specimen For The Diagnosis Of Breast Cancer,” Academic Radiology, 1995, Vol. 2, pp.199-204.

5. Lo SC, Chien M, Jong S, Li H, Freedman MT, and Mun SK: “Extraction of Rounded and Line Objects for the Improvement of Medical Image Pattern Recognition,” IEEE/MIC Proceedings, Nov. 1994 .

6. Lo SC, Lin JS, Freedman MT, and Mun SK: “Application of Artificial Neural Network to Medical Image Pattern Recongnition,” WCNN, INNS Press, 1994, Vol. I, pp.37-42.

7. Chan HP, Wei D, Helvie MA, Sahiner B, Adler DD, Goodsitt MM, Petrick N. Computer-aided classification of mammographic masses: Linear discriminant analysis in texture feature space. Physics in Medicine and Biology. 1995; 40: 857-876.

8. Wei D, Chan HP, Helvie MA, Sahiner B, Petrick N, Adler DD, Goodsitt MM. "Classification of mass and normal breast tissue on digital mammograpms: Multiresolution texture analysis." Medical Physics. 1995;22: 1501-1513.

9. Chan HP, Lo SCB, Sahiner B, Lam KL, MA Helvie. "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network." Medical Physics. 1995; 22:1555-1567.

10. Lo SCB, Chan HP, Lin JS, Li H, Freedman M, Mun SK. "Artificial convolution neural network for medical image pattern recognition." Neural Networks. 1995, Vol 8, No. 7/8, pp.1201-1214.

11. Lo SC, Li H, Krasner BH, and Mun SK, “Full-frame compression algorithms of wavelet and cosine transform,” SPIE Proc. Med. Imaging 1995, Vol. 2431, pp. 195-202.

12. Lo SC, Li H., Freedman MT, and Mun SK, “Artificial visual neural network with wavelet kernels for general disease pattern recognition,” SPIE Proceedings, Medical Imaging 1995, vol. 2434, pp. 579-588.

13. Chan HP, Wei D, Lam KL, Lo SCB, Helvie MA, Adler DD. "Computerized detection and classification of microcalcifications on mammograms." SPIE Proc. Med. Imaging 1995, Vo; 2434, pp. 612-620.

14. Sahiner S, Chan HP, Wei D, Helvie MA, Petrick N, Adler DD, Goodsitt MM: “Image classification using a convolution neural network,” SPIE Proc. Med. Imaging 1995, Vol. 2434, pp. 838-845.

15. Petrick N, Chan HP, Sahiner B, Wei D, Helvie MA, Goodsitt MM, Adler DD: “Automated detection of breast masses on digital mammograms using adaptive density-weighted contrast -enhancement filtering,” SPIE Proc. Med. Imaging 1995, Vol. 2434, pp. 590-597.

16. Wei D, Chan HP, Helvie MA, Sahiner B, Petrick N, Adler DD, Goodsitt MM: “Multiresolution texture analysis for classification of mass and normal breast tissue on digital mammograms,” SPIE Proc. Med. Imaging 1995, Vol. 2434, pp. 606-611.

17. Petrick N, Chan HP, Sahiner B, Wei D. An adaptive density weighted contrast enhancement filter for mammographic breast mass detection. IEEE Trans. Medical Imaging. 1996, Vol. 15, No. 1, pp. 59-67.

18. Lo SC, Lin JS, Li H, Hasegawa A, Freedman MT, and Mun SK, "Detection of subtle clustered microcalcifications using fuzzy modeling and convolution neural network," SPIE Proceedings, Medical Imaging on Image Processing, 1996, Vol. 2710, pp. 8-15.

19. Lo SC, Li H, Wang Y, Freedman MT, and Mun SK, "On optimization of orthonormal wavelet decomposition: Data accuracy, feature preservation, and compression," SPIE Proceedings, Medical Imaging on Image Display, 1996, Vol. 2707, pp. 201-214.

20. Osamu Tsujii, Akira Hasegawa, Chris Y. Wu, Shih-Chung B. Lo, Matthew T. Freedman, Seong K. Mun, "Classification of microcalcifications in digital mammograms for the diagnosis of breast cancer" in PROC. SPIE Proceedings, Medical Imaging on Image Processing, vol. 2710, (# 83) [Received Cum Laude Award in the Meeting]

Articles Accepted for Publication:

1. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, Goodsitt MM. "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," IEEE Trans. Medical Imaging.

2. Chan HP, Lo SCB, Niklason LT, Ikeda DM, Lam KL. "Image compression in digital mammography: Effects on computerized detection of subtle microcalcifications." Medical Physics.

Articles Submitted for Publication:

1. Li H, Liu KJ, and Lo SC, "Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms," IEEE Trans. Med. Imag.

2. Lo SC, Li H, Wang Y, Freedman MT, and Mun SK, "On optimization of wavelet decomposition for image compression and feature preservation," IEEE Trans. on Image Processing,

3. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, Goodsitt MM, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue on mammograms," Medical Physics

4. Petrick N, Chan HP, Wei D, Sahiner B, Helvie MA, Adler DD, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and tissue classification," Medical Physics

5. Wei D, Chan HP, Petrick N, Sahiner B, Helvie MA, Adler DD, Goodsitt MM, "False-positive reduction technique for detection of masses on digital mammograms: Global and local multiresolution texture analysis," Medical Physics.

Personnel Receiving Pay From This Grant

Shih-Chung B. Lo, Ph.D.
Matthew T. Freedman, M.D.
Akira Hasegawa, Ph.D.
Yuzheng C. Wu, Ph.D.
Huai Li, M.S.
Heang-Ping Chan, Ph.D.
Mark Helvie, M.D.
Nicoles Petrick, Ph.D.
Datong Wei, Ph.D.
Berkman Sahiner, Ph.D.

Pergamon

Neural Networks, Vol. 8, No. 7/8, pp. 1201–1214, 1995
Copyright © 1995 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0893–6080/95 $9.50 + .00

0893-6080(95)00061-5

1995 SPECIAL ISSUE

Artificial Convolution Neural Network for Medical Image Pattern Recognition

SHIH-CHUNG B. LO,[1] HEANG-PING CHAN,[2] JYH-SHYAN LIN,[1] HUAI LI,[1] MATTHEW T. FREEDMAN[1] AND SEONG K. MUN[1]

[1] Georgetown University Medical Center and [2] University of Michigan Medical Center

(*Received* 1 *November* 1994; *revised and accepted* 4 *May* 1995)

Abstract—*We have developed several training methods in conjunction with a convolution neural network for general medical image pattern recognition. An unconventional method of using rotation and shift invariance is also proposed to enhance the neural net performance. The structure of the artificial neural network is a simplified network structure of the neocognitron. Two-dimensional local connection as a group is the fundamental architecture for the signal propagation in the convolution neural network. Weighting coefficients of convolution kernels are formed by the neural network through backpropagated training for this artificial neural net. In addition, radiologists' reading procedure was modelled in order to instruct the artificial neural network to recognize the predefined image patterns and those of interest to experts. Our training techniques involve (a) radiologists' rating for each suspected image area, (b) backpropagation of generalized distribution, (c) trainer imposed functions, (d) shift and rotation invariance of diagnosis interpretation, and (e) consistency of clinical input data using appropriate background reduction functions.*

We have tested these methods for detecting lung nodules on chest radiographs and microcalcifications on mammograms. The performance studies have shown the potential use of this technique in a clinical environment. We also used a profile double-matching technique for initial nodule search and used a wavelet high-pass filtering technique to enhance subtle clustered microcalcifications. We set searching parameters at a highly sensitive level to identify all potential disease areas. The artificial convolution neural network acts as a final detection classifier to determine whether a disease pattern is shown on the suspected image area.

Keywords—Neural network, Computer-assisted diagnosis, Classification invariance of operations, Output association fuzzy function, Trainer imposed function.

1. INTRODUCTION

As high speed computers become cost-effective tools, many scientists have started to investigate potential technologies for computer-assisted diagnosis (Doi, 1989; Doi et al., 1992). More and more digital imaging systems are available to radiology departments as well. It is known that conventional diagnostic procedures can be enhanced by various methods through computers. The applications in computer-assisted diagnosis will be much more meaningful when clinical images are fully computerized and networks are available in radiology departments. Medical diagnoses involve very sophisticated decision-making processes. Integration of the patient information in a Picture Archiving and Communication System (PACS) (Horii et al., 1990; Huang et al., 1990) and development of computer-aided diagnosis will provide radiologists with more relevant information to significantly improve patient care.

Skilled radiologists have a high degree of accuracy in diagnosis. However, there remain problems in the detection of some diseases, problems that cannot be corrected with current methods of training and high

Acknowledgements: This project was supported in part by a U.S. Army Grant DAMD17-93-J-3007 and an American Cancer Society Grant No. EDT-93. The reviews, opinion and/or findings contained in this paper are those of the authors and should not be construed as an official Department of Army position, policy or decision unless so designated by other documentation. The LABROC program was provided by Dr. Charles Metz of the University of Chicago. The authors are grateful to the reviewers' constructive suggestion as well as to Ms Susan Kirby and Dr. Walid Tohme for their editorial assistance.

Requests for reprints should be sent to Dr Shih-Chung B. Lo, Radiology Department, Georgetown University Medical Center, 2115 Wisconsin Avenue, N.W., Suite 603, Washington, DC 20007, USA.

levels of clinical skill and experience. These problems would cause for example the miss rate in the detection of small pulmonary nodules, the detection of minimal interstitial lung disease and the detection of changes in pre-existing interstitial lung disease. In this paper, we employed a convolution neural network and proposed several training methods to enhance the detection of small pulmonary nodules and microcalcifications on digital projection X-ray images. Both diseases are clinically important in diagnostic imaging and are relatively difficult to identify when they are superimposed on other anatomical structures.

Several image processing techniques have been proposed for the detection of lung nodule: (a) thresholding and circularity calculation (Giger et al., 1988), (b) morphological operation (Giger et al., 1990), and (c) 2-D sphere profile matching technique (Lo et al., 1993). With each of these methods there is a trade-off between increased sensitivity and decreased specificity. By setting less stringent criteria with the above algorithms, the sensitivity of the detection programs would be relatively high but false detection would also increase. On the other hand, a low sensitivity setting of the program would potentially miss many true positives. To use these methods for the detection of small lung nodules, additional techniques are needed to reduce the number of false positives and maintain high true positive detection. A similar situation was found in the detection of microcalcifications in mammography (Chan, Doi & Galhotra, 1987; Chan, Doi & Vyborny 1988, 1990). For this reason, several investigators have intended to use the neural network as a classifier to improve the detection rate (Lo et al., 1993, 1996; Wu et al., 1992).

The nets of the artificial neural network used in conventional backpropagation are fully and uniformly connected from one node of the upper layer to each node in the next lower layer. When applying this type of neural network to directly perceive image patterns, the performance seems rather limited (Lo et al., 1993). In some applications the features generated by image processing techniques were used for image pattern recognition. In observing clinical radiologists' work, it is clear that they use findings in the region surrounding the suspected area when identifying the presence of a true disease. We therefore believe that the neighborhood information rather than non-local information in the image must be taken into more serious consideration during the neural network training. For direct image input, we learned that the neocognitron has been successfully used in the recognition of characters and numbers of handwriting (Fukushima, 1980, 1989; Fukushima et al., 1983; Fukushima & Wake, 1991). The neocognitron also seems likely to be able to incorporate information of the area surrounding the suspected area into its processes and has the potential to deal with ambiguity in the information set. This is the motivation to incorporate artificial visual neural network in our research for medical image pattern recognition. In this paper we propose a convolution neural network structure and several associated algorithms for general medical image applications when an abnormality of a disease pattern can be shown in a small image area. The reduction of the image area for each training or testing is recommended for two reasons: (a) a large area demands a great deal of computation and (b) it potentially defocuses the features with which the user intends to train the neural network.

2. MATERIAL AND METHODS

2.1. Fundamental Approach

Radiographs for diagnostic medical imaging have been used for many years. The diagnostic results are based on the visual pattern recognition by trained radiologists. Throughout this study we tried to mimic the radiologists' diagnostic viewing routine. Typically radiologists scan the image, looking for potential abnormalities, then evaluate each suspected area. In detecting lung nodules, radiologists first search for suspected areas on the chest radiograph, looking for bright round objects within the rib cage boundary. Next, each suspected area is examined to compare the contrast information of the bright spot to the local background. Sometimes a radiologist uses several viewing positions to look at the area. When using a workstation, the radiologist may utilize zoom and "window and level" functions to get different views about roundness and contrast information for the suspected areas. The "window" function takes a given digital value range (e.g., 3000) and rescales onto a monitor gray value range (typically 256). The "level" function selects the middle digital value for the "window". Since both window range and level can be simultaneously operated, the radiologist is able to observe various contrasts. The differentiation between a nodule and an end-on vessel can be very difficult for the human eye to discern but is often based on the presence of projections from the round shape and its relative contrast compared to the background and to other vessels. For the detection of microcalcifications, radiologists use similar viewing steps. The main differences between the detection of lung nodules and microcalcifications are the disease patterns, clinical indications, and experience.

The radiologist diagnostic viewing steps described above were modelled and were converted to computer algorithms. The detail algorithms and techniques for the pre-scan were previously described by Lo et al. (1993) for the detection of lung nodule on chest radiographs and by Chan and coworkers (1987, 1988, 1990, 1995) for the detection of microcalcifications on mammograms. In this paper, we concentrate mainly on the proposed convolution neural network and methods to adjust and arrange the input and output signals in order to achieve maximum efficiency.

2.2. The Convolution Neural Network

Based on the pre-scan methods, we set the computer programs to a highly sensitive level to extract possible objects which included all true-positive detections as well as a large number of false-positive detections. Differentiation of false positives from true positives is the remaining issue. We propose to use the trained convolution neural network (CNN) as the final classifier to carefully study each suspect area in the second phase of the diagnostic process. The proposed CNN can be considered a simplified vision machine designed to perform the second part of the disease detection study for the classification into disease and non-disease. This neural network is based on the network structure of neocognitron (Fukushima et al., 1983) which is designed to simulate the vision of vertebrate animals. We believe that the CNN should be suitable for general medical image pattern recognition.

Before entering an input matrix into the neural network, we employed a background reduction method (see Section 2.3.1) to mimic the function of "window and level". This image function has been widely used in clinical workstations. It is utilized to adjust the overall brightness of the image so that nodules of the same size would have similar contrast when compared to the background. In a way, it can help minimize the contrast variation of the disease pattern. In this study the background of all the suspected image blocks was reduced for the CNN training and testing. The purpose of using the two-dimensional convolution operation is to simulate radiologists' viewing of a suspected area. In other words, we instructed the neural net to utilize the information on both the center of the image block and its neighborhood and to train the neural network to extract necessary local features through the supervised backpropagation training. On the output side, we tried to educate the neural net by simulating the radiologists' decision making process. To model radiologists' interpretation of an image area with a certain probability of abnormality, a method of utilizing fuzzy output association is proposed in Section 2.3.3 to turn this kind of clinical measure into information readable by a neural network.

2.2.1. *The Structure of the Proposed Convolution Neural Network*. The CNN is a simplified version of the neocognitron. Since there is no theory to indicate what is the best neural network structure for medical image pattern recognition, we started our studies by using two-level and three-level neocognitron structures. We did not use complex-layer and did not extend our study beyond a three-level structure due to computation constraints. Nets between two adjacent levels (layers) are selectively interconnected across groups. The forward propagation algorithm was developed for non-supervised training in the original neocognitron method. The supervised training from one layer to the next (i.e., layer training) also proposed by Fukushima and Wake (1991) for handwriting recognition may be applied to the classification of disease patterns but will not be discussed in this paper. Instead we used a convolution constrained neural network with the well known backpropagation method for training. Figure 1 shows the global three-level structure of this neural network.

We group the operations of kernels and the image block in such a way that the center of the suspected nodule area is separated from the surrounding region. Basically each group in the receiving layer receives signals from two groups of weights (e.g., kernels). The kernels operating in the surrounding areas are referred to as peripheral kernels and the kernels operating in the central areas as inner kernels. This arrangement is specifically designed for image blocks containing a suspected tumor. In such a case, the bright spot is located relatively in the central area indicated by the pre-scan procedures. The purpose of using the dual-kernel is to instruct the peripheral and the inner kernels to learn different image patterns. However, for those tasks not involving the recognition of round objects, the use of a single kernel is recommended. In the following experiment, we use dual-kernel and single kernel for detection of lung nodules and microcalcifications, respectively. For the forward signal propagation, the resultant of the weighting factors of the kernel convoluting the element values of the front layer are collected into the corresponding matrix elements of the receiving layer. This operation accounts for the major difference between the convolution type neural network and a regular fully connected neural network. The collected value at each element is further operated with a sigmoid function in the forward propagation as it functions in an ordinary forward propagation neural network system.

Each suspected image block of 32 × 32 pixels indicated in the pre-scan program is extracted as an object for CNN classification. Due to the long training time of the computer using the CNN algorithm, every four pixels in a 2 × 2 square were averaged into one pixel so that each image block was reduced to 16 × 16 pixels. We used an array size of

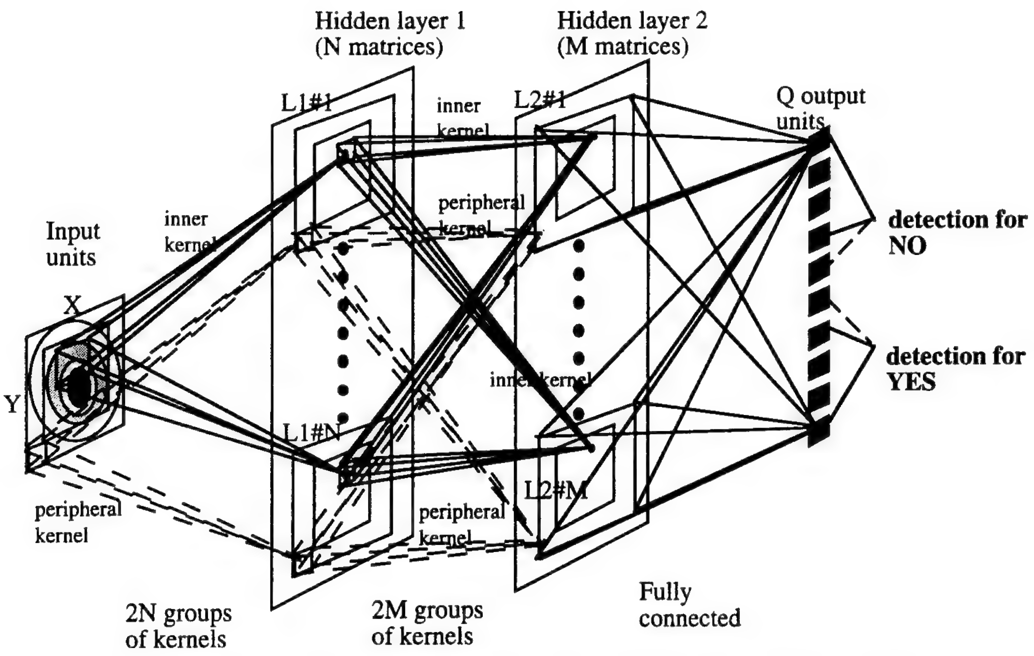


FIGURE 1. An artificial convolution neural network with dual-kernel for lung nodule detection.

5 × 5 for both inner and peripheral kernels between layers. The first hidden layer consists of 12 groups. Each group has 12 × 12 pixels formatted in a square array where the center 8 × 8 pixels and outer area covering by two pixels along the side are contributed by the inner and peripheral kernels, respectively. The second hidden layer also consists of 12 groups. Each group has 8 × 8 pixels where the center 6 × 6 area and outer area covering only one pixel along the side are contributed by the corresponding inner and peripheral kernels, respectively. The output layer has 10 nodes (groups) which are fully connected to the second hidden layer. However, in the experiment involving the detection of microcalcifications described later, no peripheral kernel was used. This is because most microcalcifications are concentrated on a few pixel regions while using a digitization pixel size of 105 μm.

It is important to realize that the total number of nodes needed in the hidden layers somewhat depends on the total number of training samples. Since we plan to expand our database and the use of rotated versions of an input matrix, we expect that our training samples will be very large in the future. The number of layers used should depend upon the sophistication of the features that the neural network is intended to perceive. The more complicated the disease patterns, the more layers are required to distinguish high order information of image structures. The convolution kernels are organized in such a way as to emphasize a number of image characteristics rather than those less correlated values obtained from feature spaces for input. These characteristics are: (a) the horizontal versus vertical information; (b) local versus non-local information; and (c) image processing (filtering) versus signal propagation.

2.3. Image Processing and Training Methods

An appropriate neural network structure is an important working base to form a signal propagation platform in a given recognition task. The training materials and methods, which provide intellectual information for the construction of the knowledge, are essential for the performance of the neural network. We believe that the success of using the neural network relies not only on the network structure but also on the sufficient training information and effective training methods. This study demonstrates our approaches to convert expert knowledge into computer readable information, which is the key issue in terms of training. In this experiment, we provided the network with all possible radiological diagnostic information and set up the studies by adding one method at a time to optimize the neural network performance.

2.3.1. *Background Reduction for Suspected Image Blocks.* We found that the consistency of input matrix contrast is an important factor in stabilizing the neural network learning. In our experiment, the neural network did not reach a solution for the image blocks provided in the training sets, even though all suspected nodule blocks were corrected for background trend and relatively centered in terms of brightness. Their contrasts (the difference between the center and peripheral brightness) are unevenly distributed due to a variation of X-ray exposure and different sensitivity of the films. In addition, the image block may involve many superimposed background structures, namely vessels, ribs, and the heart. Separating nodules or suspected round objects from chest structures is not an easy task. We concluded that elimination of some background information and enhancement of contrast information are necessary procedures to assist the neural network in the recognition of disease patterns. We designed a background subtraction technique to simulate "window and level" function, which is clinically useful for enhancing disease patterns. In fact, we only used "level" function and ignored "window" function. A fixed "window" function may distort and mix the contrast information that exists between nodules and end-on vessels. For the "level" function, gray values in each image block are uniformly subtracted from the calculated background by averaging the outer ring area.

$$fs(x,y) = \begin{cases} f(x,y) - B & \\ \quad \text{for } f(x,y) > B \text{ and } (x,y) \text{ is inside the circle} & \\ 0 & \\ \quad \text{for } f(x,y) < B \textit{ or } (x,y) & \\ \qquad \text{is on or outside the circle,} & \end{cases} \tag{1}$$

where

$$B = \frac{\sum_{n \in \text{the ring}} f(x_n, y_n)}{C}$$

for circular object detection and C equals the number of pixels in the ring. Figure 2 shows that heavily-shaded pixels are used in the calculation for background averaging. Both heavily and lightly shaded areas are given in pixel value of 0. Our studies indicated that this ring area averaging method produces better results than the peripheral area averaging method. This may be due to the fact that the ring area is closer to the central area and possesses greater background information than the entire peripheral area.

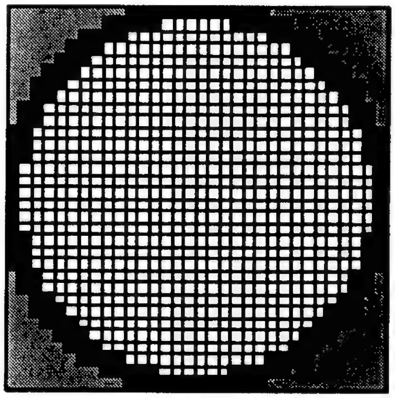


FIGURE 2. A 32 × 32 image block. The white area is the area of interest for image pattern recognition using convolution neural network. Original pixel values in the heavily shaded area are averaged as a background value.

For the detection of non-circular objects, the background value should be obtained by averaging pixel values on the frame. Each gray value from the background-reduced image block is one-to-one transferred to a node of the input layer for the neural net processing. These signals received at the input layer are equivalent to the light signals received by the retina as far as the vision type neural network is concerned.

2.3.2. *Backpropagation Training*. The main difference between conventional weights and kernel weights is that conventional weights are independent and kernel weights are constrained by grouping. We believe that the latter method is more powerful than the former method for direct image pattern recognition. In addition, the trained kernels can be analyzed to understand what features were learned during the training. This design would allow researchers to further investigate the artificial neural network learning. Training requires many iterations for the network to obtain solutions for all weights applied to the propagation while the error function reaches a minimum value.

By looking at the CNN processing, one may find that signals are filtered and modulated as in a complicated circuit system. Signal propagation from one layer to the next is composed of a two-step calculation: (a) adaptive convolution combiner and (b) an activation function (a sigmoid function is used in this study) which is given below:

$$S_x((i,j);n) = \frac{1}{1 + \exp\left\{ - \sum_m [k_x((u,v);n,m)) \otimes S_{x-1}((i,j);m))] \right\}} \tag{2}$$

or

$$S_x((i,j);n) = \frac{1}{1 + \exp\left\{ - \sum_{u,\,v;\,m} [k_x((u,v);n,m)) \times S_{x-1}((i-u,j-v);m))] \right\}}, \tag{3}$$

where $S_x((i,j);n)$ represents the signal at node (i, j),

nth group, and x layer; $k_x((u,v);n,m)$ denotes the weighting factor value of net (u, v) in the nth group of the $x - l$ layer which connects the mth group of the x layer.

The error function which is expected to reach a local minimum through the error backpropagation training can be given as:

$$E = \frac{1}{2}\sum_{n_o=1}^{T}[y(n_o) - S_o(n_o)]^2, \qquad (4)$$

where $y(n_o)$ and $S_o(n_o)$ are the target output and calculated output signals for output node n_o, respectively and T is the total number of output nodes. Based on eqns (3) and (4), the iterative version of kernel weights derived by the generalized delta rule is given as:

$$\begin{aligned} &k_x((u,v);n,m)[t+1] \\ &\quad = k_x((u,v);n,m)[t] \\ &\qquad + \eta\sum_{i,j}\delta_x((i,j);n)S_{x-1}((i-u,j-v);m) \\ &\qquad + \alpha\Delta k_x((u,v);n,m)[t], \qquad (5) \end{aligned}$$

where t is the iteration number during the training, η is the gain for the current weight changes, α is the gain for the momentum term received in the last learning loop, and δ is the weight-update function which is given as:

$$\delta_x(i,j);n) = S_x((i,j);n)[1 - S_x(i,j);n)]Q_x((i,j);n) \qquad (6)$$

and

$$Q_x((i,j);n) = \sum_{u,\,v;\,m} k_{x+1}((u,v);n,m) \times \delta_{x+1}((i+u,j+v);m).$$

For the output layer,

$$\begin{aligned} \delta_o((i,j);n_o) &= \frac{\partial E}{\partial S_o(n_o)}S_o(n_o)[1 - S_o(n_o)] \\ &= [S_o(n_o) - y(n_o)]S_o(n_o)[1 - S_o(n_o)] \qquad (7) \end{aligned}$$

where o denotes the output layer. In this study, all weighting factors including the kernels were initially given a normalized random number. The normalization is based on the number of nets connecting to a destination node in the next layer.

2.3.3. *Neural Network Output Assignment Using Radiological Diagnostic Rating*. The design of the output layer for the medical diagnostic decision is not as straightforward as in other applications. Our goal is to distinguish non-disease patterns from disease patterns. We can classify the data set in two categories. However, this is probably not an optimal design for the output layer. In some obvious cases, radiologists are able to make a clear diagnostic indication of a disease shown on an image. Often they work with different degrees of sensitivity (different levels of suspicion) depending on the clinical situation. Thus they estimate the likelihood that a radiograph or an area of a radiograph may possess a disease. For the neural network, it may be more realistic to define the output in terms of probability. Depending upon the number of output nodes used, the arrangement of output nodes and the probability associated with a score varies. Intuitively, one can proportionally scale scores onto node numbers.

Although the above output node assignment follows the general diagnostic decision rule used by many radiologists, one output node has no relation to any other. No output node relation will be passed to the neural network for the training. To circumvent this problem, we propose to use a narrow output distribution to establish a fuzzy association between the adjacent output nodes. In fact, when a radiologist determines a specific probability of a disease pattern in an image area based on his/her training and experience, this probability would be accompanied by a variation. We modelled this probability with a generalized distribution (Szepanski, 1980) in the score space.

$$G(\sigma, v, p) = Ce^{-|gv|^p} \qquad (8)$$

where v is the distance from a given score,

$$C = \frac{pg}{2\Gamma\left(\frac{1}{p}\right)} \text{ and } g = \frac{\sqrt{\Gamma\left(\frac{3}{p}\right)}}{\sigma\sqrt{\Gamma\left(\frac{1}{p}\right)}}$$

The reason for modelling a generalized distribution is that we do not know exactly what kind of distribution can represent the radiologists' interpretation in various diseases. When $p > 2$, the distributions may be too flat which probably is not the case with highly experienced pulmonary radiologists. For simplicity, we use $p > 2$ for Gaussian distribution in the experiment.

In addition to the distribution function, the trainer can impose a driving function, $r(v, s)$, onto it to indicate the belonging of the determination category, where s denotes the strength of the repulsion introduced by the user. An example of the trainer

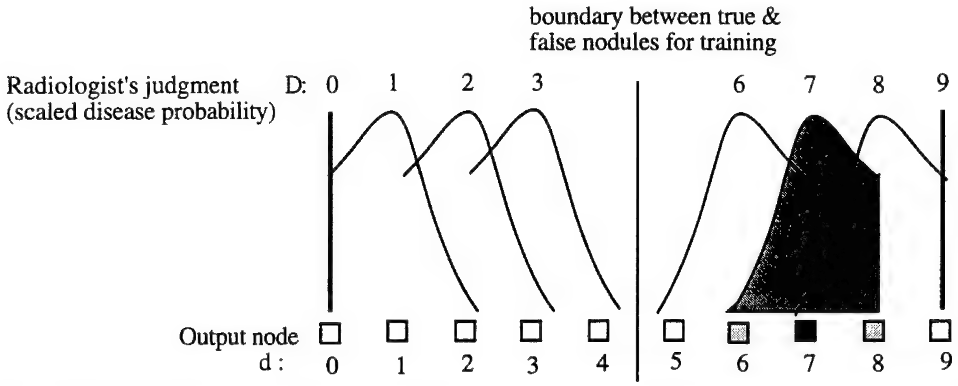


FIGURE 3. A fuzzy output association is constructed by Gaussian distribution and repulsive functions. (Note this drawing is not in scale. Only one curve is used for a training case.)

imposed driving function for scores indicating positive determination is given below:

$$r(v,s) = \begin{cases} 1 & \text{for } v \leqslant 0 \\ sv+1 & \text{for } v > 0. \end{cases} \tag{9}$$

For those scores associated with negative determination, the repulsion of eqn (9) should be changed to the opposite direction. Therefore, the output association functions at a single node for a score indicating an image block involving disease and for a score indicating a disease-free image block are:

$$Ah = K \times G(\sigma, v, p) \times r(v,s) \tag{10}$$

and

$$Al = K \times G(\sigma, v, p) \times r(-v,s), \tag{11}$$

respectively. For the extreme scores in the score space (i.e., minimum and maximum scores), the use of a delta function is recommended.

In the lung nodule detection studies, we assigned scores for all image blocks for training. Based on the score which corresponds to an output node, a Gaussian distribution ($p = 2$) with a standard deviation of $\sigma = 0.55$, an output scaling constant of $K = 2.5$ and a repulsive strength of $s = 1.5$ for the asymmetric output association ($s = 0$ for the symmetric output association) were used for correlating the adjacent scores. For the neural network output, we used a discrete form of the score. We estimate that it would take a great deal of computation time for 100 nodes or more in the output layer. Realistically, 10 discrete output nodes are proposed for the classification. We assigned nodes 0–3 to correspond to definitely negative-possibly negative; detection nodes 6–9 correspond to possibly positive-definitely positive detection. Nodes 4 and 5 are not used for decision buffering. During the experiment, we collected all the suspected nodes in two categories (i.e., true nodule and non-nodule). In the course of rating for the training set, a senior radiologist scored each suspected nodule based on his clinical knowledge. Pathologically proven truth (either has a nodule or not) of training case was also provided to assist in the radiologist's rating.

Figure 3 shows all the asymmetric output association distributions corresponding to a radiologist's judgement. However, only one curve was used for each judgement with a suspected image block. Figure 3 also highlights a case when score 7 is determined. In this situation, output node 7 received the highest activation (1.0), node 8 received the second highest activation (0.5), node 6 receives some activation (0.2), and remaining nodes receive no activation.

Two examples of output assignments associated with probability in discrete form are given below:

(a) Symmetric output assignment:

$$A(d,D) = \begin{cases} 0.2 & \text{for } 9 > D > 5 \text{ and } d = D+1 \\ & \quad \text{or for } 0 < D < 4 \text{ and } d = D-1 \\ 1.0 & \text{for } D = d \\ 0.2 & \text{for } 9 \geqslant D > 5 \text{ and } d = D-1 \\ & \quad \text{or for } 0 \leqslant D < 4 \text{ and } d = D+1 \end{cases} \tag{12}$$

otherwise $A(d,D) = 0$.

(b) Asymmetric output assignment:

$$A(d,D) = \begin{cases} 0.5 & \text{for } 9 > D > 5 \text{ and } d = D+1 \\ & \quad \text{or for } 0 < D < 4 \text{ and } d = D-1 \\ 1.0 & \text{for } D = d \\ 0.2 & \text{for } 9 \geqslant D > 5 \text{ and } d = D-1 \\ & \quad \text{or for } 0 \leqslant D < 4 \text{ and } d = D+1 \end{cases} \tag{13}$$

otherwise $A(d,D) = 0$.

The use of asymmetric output assignments attempted to push (train) the non-disease pattern toward low score nodes and to instruct the disease pattern toward high score nodes. With this output assignment for the output node in the training, the adjacent node relation is also established. This supervised training can be generally applied to any situation where association of outputs is necessary.

2.4. Classification Invariance of Matrix Operations

The use of moment invariance via rotation and shift has been proposed for applications in graphic pattern recognition. The direct use of this method as a classifier may not be suitable for those image patterns possessing circular symmetric property (e.g., nodule) or lacking a fixed geometric pattern (e.g., calcification).

Often medical image pattern recognition does not concern "top-down" or "left-right" as classification criteria. In such a case we can take advantage of this characteristic as an invariance. In other words, we propose to rotate and/or to shift the input vector and maintain the same output assignments for the training. This method may affect the neural network in two ways: (a) by instructing the neural network that the rotation and shift of the input vector would receive the same classification result; and (b) by increasing the total number of training samples which is expected to enhance the performance of the neural network.

Using the center pixel as the origin, a standard rotation and shift of the image block was used;

$$\begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix} = \begin{bmatrix}\cos(\phi) & \sin(\phi)\\ -\sin(\phi) & \cos(\phi)\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} + \begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}, \tag{14}$$

where ϕ is rotation angle of the origin (center of a 32 $\times$ 32 image block); Δx and Δy are shifts in the x and y directions, respectively.

In this work, we only rotated each input matrix eight times to test our hypothesis. Four of the rotations are:

$$\phi \in \{0°, 90°, 180°, 270°\}. \tag{15}$$

We also flipped over (left-right) the original image matrix and used the above rotations again to obtain four additional rotations. This type of rotation would only reposition pixel values. No interpolation calculation of pixel values was involved. We believe that other rotations and minor shifts are also valid methods for the use of classification invariance of operations. Rotation may require interpolation which would slightly alter the pixel values and should be acceptable for the input of the CNN. However, the use of shifting can be complicated, because it involves (a) how important the center information for disease patterns are in the neural network learning and (b) how much shifting can be used without sacrificing critical portions of image information.

2.5. Classification of Output Values in the Testing

We assigned scores with a narrow asymmetric peak distribution on the output nodes for the training in order to associate each node with its adjacent node. We believe that the distribution assignment is not a unique method to link rating score relations. However, the output relation information must be passed to the neural network for learning. This relationship does not exist in recognition for characters or Arabic numbers. In those applications, each node is independent from others.

After the training a typical output pattern will be very close to the corresponding perfect pattern (the assigned narrow asymmetric peak distribution) for most of the training cases. In the case of testing, many of them have different output signal patterns. It is not a simple task to interpret what the representation of each output pattern means if the testing output does not follow an output pattern assigned to the training. Corresponding to the grading system arranged in the training, a polarized (linearly weighted) function is given as an indication. With this we can define a normalized disease detection index (NDDI) for the judgement of a suspected area:

$$\text{NDDI} = \frac{\sum_{n=N/2}^{N-1} [O_n \times (n - (N-1)/2)]}{\sum_{n=0}^{N-1} [O_n] \times (N-1)/2} \tag{16}$$

where n denotes the node in the output layer, O_n is the output value at node n, and N is the total number of output nodes. Hence a nodule detection index of 0 or near 0 indicates a definite non-nodule and a nodule detection index of 1 or greater implies a definite nodule case with the judgement of the neural network. The reason for the weighting is that the score line is centered at $(N-1)/2$ (i.e., 4.5 for 10 nodes in the output layer) and polarization of true and false depends on the position of the nodes. Equation (16) is the net effect of all output nodes.

We do not recommend using a detection index of 0 as a cut-off point to determine a disease or a disease-free image block using the trained neural network. The cut-off point may be shifted by the inevitable bias in the training cases. In practice the cut-off point is established by many clinical cases in a rigorous evaluation study. In this paper a pre-clinical performance study was conducted and is discussed below.

2.6. Performance Evaluation of The Convolution Neural Network

Receiver operating characteristic (ROC) is an analytical method generally applied to the performance evaluation of a system. In an ROC analysis, the distributions of the normal and abnormal cases may be represented by binormal distributions (Swets & Pickett, 1982). When the two distributions overlap on the decision axis, a cut-off point can be made at an arbitrary decision threshold. The corresponding true-positive fraction (TPF) versus false-positive fraction (FPF) for each threshold can be indicated in Cartesian coordinates. By marking several points on the plot, curve fitting can be employed to construct an ROC curve. The area under the curve referred to as *Az* can be read as a performance index of the system using ROC analysis. In general the higher the *Az*, the better the performance. A computer program (LABROC) using two sets of data, one for true and the other one for false categories, is employed for the analysis of the NDDIs derived from neural net outputs.

3. EXPERIMENTS AND RESULTS

3.1. Detection of Lung Nodules on Digital Chest Radiographs

Chest radiographs in patients with primary and metastatic cancer and with one or several lung nodules are converted into digital form using a laser film digitizer (Konica Laser Film Scanner Model: KDFR-S; Tokyo, Japan). About one third of chest images were acquired from a computed radiographic system (AGFA ADC prototype computed radiography; Mortsel, Belgium). The digital data are transmitted and stored in our PACS until needed for the research project. The images were then retrieved to a high speed workstation and the computer searches were used sequentially: a thresholding evaluation, use of background reduction, a test of profile matching rate, and neural network classification.

The pre-scan process was performed first to locate the center of the island and isolate the image block for training. The pre-scan program was running in a highly sensitive mode with a matching rate (MR) of 0.7 for all images involved in the training. Suspected image blocks included various types of rib crossing, and various sizes of end-on vessels and vessel clusters. The true-positive nodules may also overlap with lung, vessels, and rib structures. Figure 4 shows some randomly sampled suspected image blocks which were background-reduced and contrast-balanced for display purposes. These image blocks were mirrored and rotated 90°, 180°, and 270° for the training. Note that each original and its seven "brother" image blocks share the same score vector (probability of a disease and output fuzzy association). During the training, the original and its seven "brother" image blocks as a group were entered in the same sequence.

During the training we found that the error-function did not monotonously decrease for each learning epoch. However, the overall errors decreased throughout many iterations. We did not completely

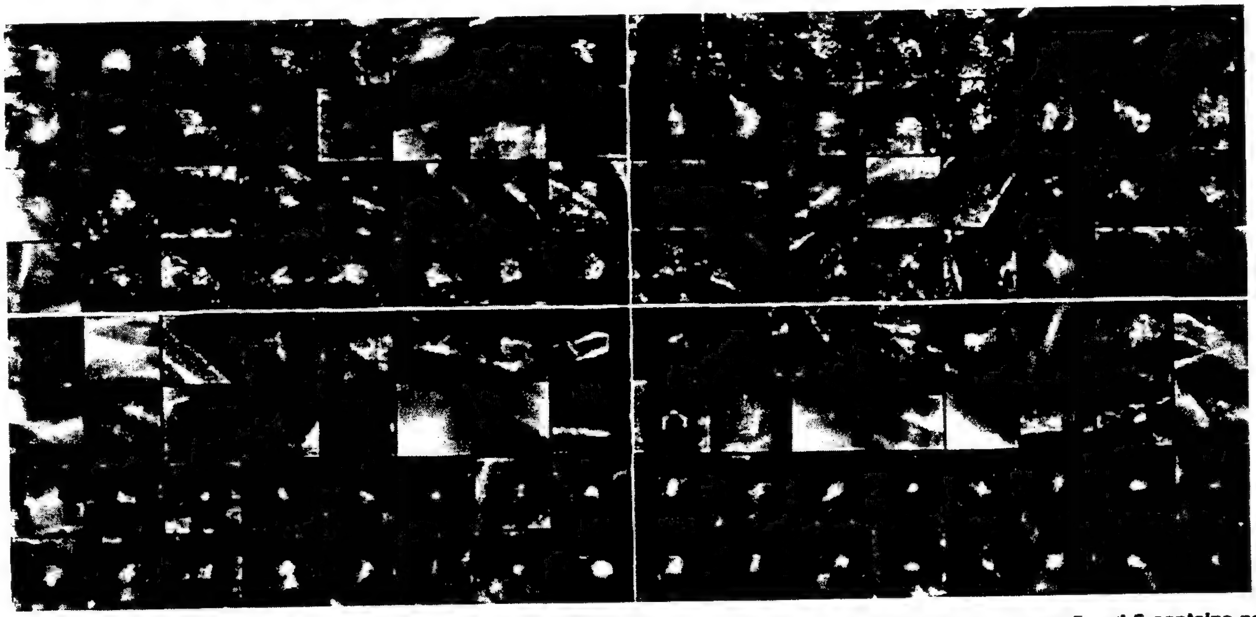


FIGURE 4. The upper four rows show 64 nodule blocks sampled from the database. Each image block on rows 5 and 6 contains no nodule but lung or rib structure. Each image block on the bottom two rows contains an end-on vessel.

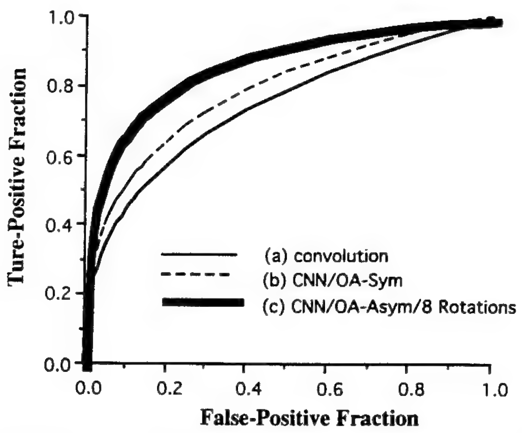


FIGURE 5. Three ROC curves representing the performance of (a) CNN without output association and rotation (plain line, the average $Az = 0.77$), (b) CNN using a Gaussian output association in the training (dashed line, the average $Az = 0.83$), and (c) CNN with asymmetric output association and eight rotated input matrices (bold line, the average $Az = 0.88$).

retrain the kernels for different assignments in the output layer. Instead, based on the trained kernels we continued to train the neural network with additional conditions. The sequence is: (a) CNN with symmetric output association, (b) use of trainer imposed driving function, and (c) rendering seven "brother" images for training.

The database had 55 chest radiographs and only 25 images contained at least one nodule. In the prescan, 52 nodules and 155 non-nodules were extracted from all 55 images. All cases were confirmed by biopsy or by follow-up showing growth of the nodule. In this study, we employed a grouped jackknife method (Fukunaga & Hayes, 1989) to evaluate the performance of the CNN. We randomly selected 28 images for training and the other 27 images for testing in the study. Final ROC curves were obtained by averaging the results from 30 grouped jackknife experiments. The results obtained from the tests were very encouraging. Figure 5 shows the improvement of using a convolution neural network and corresponding enhancement techniques using output association and classification invariance of matrix operation for the input.

In this experiment, we found that the average *Az* was 0.77 using the CNN with a delta function for output determination, and was 0.83 using the CNN with a narrow Gaussian distribution for output association. Using a Gaussian output association and eight types of rotated image blocks for input, we found that the *Az* was increased to 0.87. After a trainer imposed function was added, we obtained an insignificant increase of *Az* to 0.88. From the ROC curve corresponding to *Az* = 0.88, we found that the CNN reduced 79% of false-positive detections equivalent to 2–3 false nodule detections per image and preserved 80% of true-positive detections.

We also tested the same database using two nodes in the output layer. In such a case, no output association can be used. The CNN achieved an average *Az* of 0.83 when eight input matrices shared the same diagnostic interpretation (true or false).

3.2. Detection of Microcalcifications on Digital Mammograms

We also evaluated the use of CNN in the detection of subtle microcalcifications. A total of 68 mammograms (only 38 of them consisted of subtle

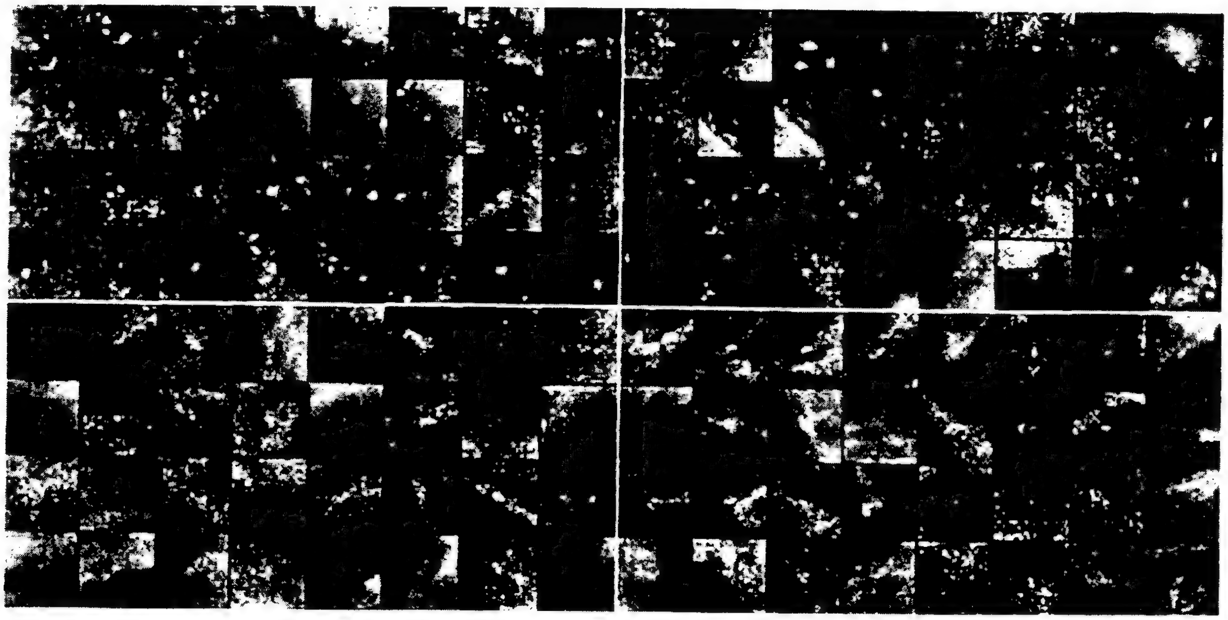


FIGURE 6. Each image block, extracted from the mammogram, on the upper four rows contains at least one calcification. Each image block on the bottom four rows contains at least a local maximum value of gray scale (bright spot) that is not a calcification. Each block at matrix elements (1,4), (5,4), 7,4), (9,4), and (2,6) contains a bright spot due to a film defect.

microcalcifications) were digitized by a laser scanner with a pixel size of 0.105 mm. The initial search prior to the final interpretation by the neural network follows the basic scheme which uses background removal and signal extraction methods to pre-scan the mammograms and to extract all possible suspected areas (Chan et al., 1988, 1990, 1991). After the pre-scan process by the computer program, the 68 digital mammograms provide 265 true and 1821 false subtle microcalcifications. Figure 6 shows some of the suspected regions which may or may not contain microcalcifications.

Prior to the CNN process, the background of all the image blocks were removed using a wavelet high-pass filtering technique instead of using the circular averaging method described in Section 2.3.1 where lung nodule detection was the objective. Specifically, after extracting each suspected region from the original digital mammogram, a three-level wavelet transform was used and only the lowest frequency was eliminated for high-pass filtering before image reconstruction. The high-pass filtered image blocks were used as the input of the CNN. For this study, we also employed the grouped jackknife method to evaluate the performance of the CNN. We did not ask radiologists to rate image blocks in the mammography training set. Only two output nodes with eight rotations for input were used. Neither output association nor trainer imposed function was employed.

In the first study, we randomly selected two sets of mammograms (i.e., 34 for training and 34 for testing) with variable sizes of kernels and image block. Figure 7 shows the *Az*s of various CNN structures used in the experiment with the same data set described above. In this figure, the CNN structures are indicated by *Sn*/*Hm*/*Kt* representing $n \times n$ pixels for input, m hidden layers, and with a kernel size of $t \times t$. The image blocks are centered on a suspected calcification indicated by the pre-scan method. This study indicated that significantly higher *Az*s were obtained when a square area of 1.7 mm (i.e., 16 × 16 pixels) region for the input and kernel size of 0.52 mm (i.e., 5 × 5 pixels) were used. In addition, the use of two hidden layers is better than the use of one hidden layer. We also found that the best results are obtained at a relatively large square error (i.e., cost function was 40–70 for 2104 cases) which suggests a fuzzy membership in the output or that more nodes in the output layer may be necessary for the optimization of the CNN in the detection of this database.

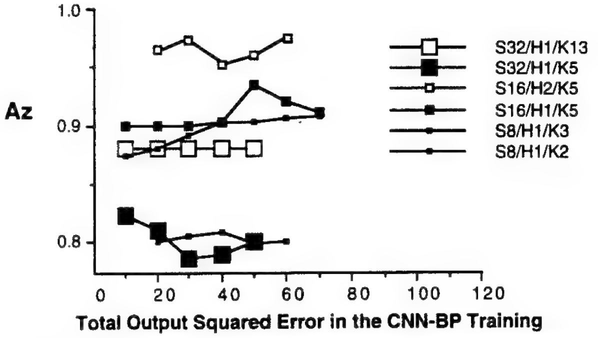


FIGURE 7. *Az*s in the detection of clustered microcalcifications using different CNN parameters.

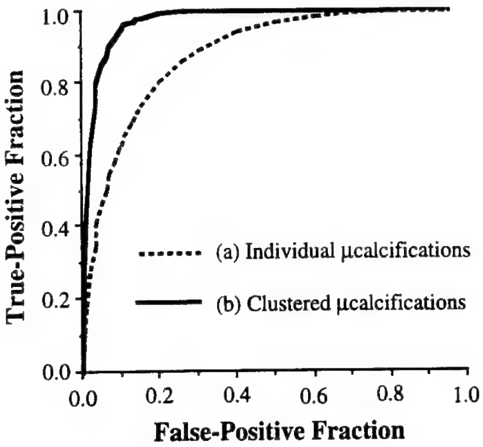


FIGURE 8. Two ROC curves representing the performance of (a) CNN using two outputs and eight types of rotation for input with the determination based on individual microcalcifications: the average *Az* = 0.89 and (b) CNN using two outputs and eight types of rotation for input with the determination based on clustered microcalcifications: the average *Az* = 0.97.

Based on the above initial study, we decided to use the CNN structure with the parameter of S16/H2/K5 for the grouped jackknife study of the CNN performance. Final ROC curves were obtained by averaging the results from 30 grouped jackknife experiments. Figure 8 shows the results of using the CNN and classification invariance of matrix operation for the input. In this experiment, the average *Az* was 0.89 when the determination was based on individual microcalcifications and was improved to 0.97 when the determination was based on the clustered microcalcifications. In the latter method, suspected clusters including one or two calcifications were rejected and the average NDDI taken from the clustered calcifications was used for the ROC evaluation. One must realize that the detection of clustered microcalcifications is more clinically significant than individual calcifications, since the clustered microcalcifications (three or more) are a strong indication of breast carcinoma in radiological diagnosis. The clustering procedure was done by grouping the detected microcalcifications in a 1 cm^2 region of the mammogram. Only a minimum of three clustered microcalcifications was considered a detection. The average ROC curve for the detection of clustered microcalcifications indicated that the CNN can eliminate 90% of false-positive detections, resulting in 0.5 false clustered detection per image, and preserve a true-positive detection rate of 87%.

4. DISCUSSION

Medical image pattern recognition using feature extraction as an input has been proposed in the detection of disease patterns (Wu et al., 1991). Since only a small number of inputs are used (as compared to 16 × 16 input signals), less computation is necessary for training. As long as the features of a disease pattern are well defined and can be quantified as values or vectors, a nonconvolution neural network should be able to classify the features. However, both the proposed diagnosis invariance and the output assignment methods for the enhancement of disease detections may only be used in limited cases. On the other hand, the structure of the convolution neural network is complicated and requires more computations, particularly for the training. The CNN does not require the feature extraction of disease patterns from the image and is capable of distinguishing non-disease patterns from disease patterns. A potential advantage of using the proposed CNN is that feature extraction can be more specifically defined not only by the user's experience but also by the confirmation of the CNN when the function of each kernel is discovered. Some complementary features learned by the CNN may be able to contribute image information of a disease thereby assisting the radiologist in better understanding all the features of the disease. Further investigation of the CNN specified features, other than known features, should be very interesting to radiologists and imaging scientists.

In this work we used preliminary scanning methods to define suspected abnormal areas. The final disease classification was analyzed by using an artificial convolution neural network with backpropagation training. We proposed several methods to mimic the radiologists' reading patterns in detecting diseases on radiographs. Though conventional image processing techniques can capture true diseases, many false-positive detections are obtained. We found that the CNN substantially reduced the number of false-positive detections.

In this study, we designed the convolution neural network to focus on local information with expert-trained output distribution. The use of diagnosis invariance of rotation seems likely to enhance the performance of the CNN by virtually increasing the number of training cases. It is obvious that both the expert-trained output distribution and the classification invariance of matrix operations are not only applicable to CNN but also to a conventional neural network as long as an image (or an image associated vector which depends on image orientation) is used in the input layer.

Summarizing the failure cases in the study of lung nodule detection, we found that the majority of false-negatives related to nodules partially overlapped with rib and many false-positives related to end-on vessels. This is because our training database was small and did not have enough true cases to cover various situations in rib overlapping on nodules and did not have enough false cases to cover various contrasts of end-on vessels. We believe that the performance of the CNN will be greatly improved when the training cases are sufficiently expanded in the future study.

One may interpret eqn (16) as another network fully connected to a single output node. This subnetwork can be included in the backpropagation training with a linear activation function for the output node. However, this subnetwork does not ensure that the backpropagated signals on the previous layer (i.e., the output layer consisting of 10 nodes) are matched with the radiologists' scores. The use of 10 nodes in the output layer also provides flexibility for the researcher to investigate the migration of kernel changes when an additional training strategy is added. The fuzzification of the teaching signals and the use of a trainer imposed function are examples of the training strategies used in this paper. The kernel changes corresponding to the training can be important information for future optimization of the CNN algorithm associated with disease pattern recognition.

In this study, we learned that the background reduction was a necessary procedure for the detection of both lung nodules and mammographic microcalcifications otherwise the error function would not reach a minimum for the training data set. Several broad output distributions were also tested. The CNN performance (i.e., generation) of those tests were inferior to that of the narrow output distribution. A comparison experiment was also conducted to evaluate the difference between the training using image groups (the original and its seven "brother" image blocks as one group) and image blocks (all image blocks). We found that the CNN seems to perform better using image groups than randomizing each image block in the training. We also modified our neural network structure to one hidden layer. The CNN performance with one hidden layer was not as effective (the average $Az = 0.81$ for kernel size of 5 × 5 and the average $Az = 0.85$ for kernel size of 13 × 13) as when two layers were used for the detection of microcalcifications. However, the performance was about the same with one hidden layer and two hidden layers for the study involving lung nodules. We do not know whether this effect was due to the fine structure of microcalcifications or smaller samples used in the experiment of lung nodule detection. We are currently testing three hidden layers to see if there is any improvement in the generalization. We are also working on 32 × 32 original image block and expect better outcomes. By increasing the sizes of image block and kernel, the computation time will increase 10–16 times as much as the training time needed for the CNN configuration described in Section 2.2.1. When the database is expanded, a higher power computer will be required for the training.

5. CONCLUSIONS

In this study, we have added several effective techniques to the convolution neural network for the enhancement of disease diagnosis: (a) development of a better background reduction method so that the neural network has a better "observation" of the image block, (b) providing radiologists' rating scale for the backpropagation training, (c) introducing the neural network with the classification invariance of input matrix operations, (d) use of output association functions to mimic the radiologists' interpretation and to establish the relationship between adjacent output nodes, and (e) rendering trainer imposed functions to enhance the performance of CNN. We found that the performance of the CNN in detecting disease was improved significantly by administering these training methods.

Studies in the use of chest radiographs for the detection of lung nodules (Stitik et al., 1985; Hellan et al., 1984) have demonstrated that even with highly skilled and highly motivated radiologists working with high quality chest radiographs, only 68% of all retrospectively detected lung cancers were detected prospectively when read by one reader, and only 82% were detected by two readers. Our studies did not have the same clinical setting as Stitik and Hellan's due to our smaller database, therefore, we could not compare our results with the radiologists' sensitivity of 68% mentioned above. However, we consider it likely that radiologists will benefit from the use of a nodule detection program such as this in one of two ways. First, the radiologist will use the program as a second reader, thus increasing the detection of lung nodules similar to the results seen in the study by Stitik. In the second method, the radiologist may call on the system as a consultant on an individual suspected area. The radiologist can point to the suspected area and ask for the interpretation from the CNN system. The CNN system, in fact, may be able to work as a trained referral system for the consultation of detecting lung nodules. Such a program is also readily available in our computer and clinical evaluation is in progress. A fully automatic lung nodule detection program takes 12–18 s for a 512 × 512 digital chest radiograph in a DEC Alpha workstation. To evaluate an identified area, it only takes the CNN program 0.2 s to respond.

This work has demonstrated two successful medical diagnostic applications using an artificial visual neural network and expert-trained computer procedures instead of a non-convolution neural network or other conventional classification method. This technique attempted to simulate the radiologists' reading pattern: pre-screen and classification for interpretation. We believe that the proposed convolution neural network and its associated training techniques can be extended to many diagnostic imaging areas such as the detection of low contrast mass in mammography and the pattern recognition of interstitial lung disease in chest radiography. In fact, the proposed CNN technique should be able to be trained to detect almost all disease patterns perceivable by a trained radiologist.

REFERENCES

Chan, H. P., Doi, K., & Galhotra, S. (1987). Image feature analysis and computer-aided diagnosis in digital radiography: 1. Automated detection of microcalcifications in mammography. *Medical Physics*, **14**, 538–548.

Chan, H. P., Doi, K., & Vyborny, C. J. (1988). Computer-aided detection of microcalcifications in mammograms. *Investigative Radiology*, **23**, 664–671.

Chan, H. P., Doi, K., & Vyborny, C. J. (1990). Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer aided diagnosis. *Investigative Radiology*, **25**, 1102–1110.

Chan, H. P., Lo, S. C., Sahiner, B., Lam, K. L., & Helvie, M. A. (1995). Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network. *Medical Physics* (in press).

Doi, K. (1989). Feasibility of computer-aided diagnosis in digital radiography. *Japanese Journal of Radiological Technology*, **45**, 653–663.

Doi, K., Giger, M. L., & MacMahon, H. (1992). Potential usefulness of real-time computer output to radiologists' interpretations. Scientific Exhibit, Space 10-001. Presented at RSNA 1992, Chicago, Ill.

Fukunaga, K., & Hayes, R. R. (1989). Effects of sample size in classifier design. *IEEE Transactions on Pattern Analysis of Machine Intelligence*, **PAMI-11**, 873–885.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, **36**, 193–202.

Fukushima, K. (1989). Analysis of the process of visual pattern recognition by the neocognitron. *Neural Networks*, **2**, 413–420.

Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, **13**(5), 826–834.

Fukushima, K., & Wake, N. (1991). Handwritten alphanumeric character recognition by the neocognitron. *IEEE Transactions on Neural Networks*, **2**, 355–365.

Giger, M. L., Doi, K., & MacMahon, H. (1988). Image feature analysis and computer-aided diagnosis in digital radiography: 3. Automated detection of nodules in peripheral lung field. *Medical Physics*, **15**, 158–166.

Giger, M. L., Ahn, N., & Doi, K. (1990). Computerized detection of pulmonary nodules in digital chest images: Use of morphological filters in reducing false-positive detections. *Medical Physics*, **17**, 861–865.

Hellan, R. T., Flechinger, B. J., & Melamed, M. R. (1984). Non small cell lung cancer: Results of the New York screening program. *Radiology*, **151**, 289–293.

Horii, S. C., Mun, S. K., & Levine, B. A. (1990). PACS clinical experience at Georgetown University. *Computerized Medical Imaging and Graphics*, **15**(3), 183–190.

Huang, H. K., Kangarloo, H., & Cho, P. S. (1990). Planning a total digital radiology department. *American Journal of Radiology*, **54**, 635–639.

Lo, S. B., Freedman, M. T., & Lin, J. (1993). Automatic lung nodule detection using profile matching and back-propagation neural network techniques. *Journal of Digital Imaging*, **6**(1), 48–54.

Lo, S. B., Lou, S. L., & Lin, J. (1996). Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transaction on Medical Imaging* (in press).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representation by error propagation. In D. E. Rumelhart & J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing*, (Vol. 1, pp. 318–362). Cambridge, MA, MIT Press.

Stitik, F. P., Tockman, M. S., & Khouri, N. F. (1985). Chest radiology. In A. B. Miller (Ed.), *Screening for cancer*, (pp. 163–191). New York: Academic Press.

Swets, J. A. & Pickett, P. M. (1982). *Evaluation of diagnostic systems*. New York: Academic Press.

Szepanski, W. (1980). Δ-entropy and rate-distortion bounds for generalized-Gaussian information source and their applications to image signals. *Electronics Letters*, **16**(3).

Wu, Y., Doi, K., Giger, M. L., & Nishikawa, R. M. (1992). Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks. *Medical Physics*, **19**, 555–560.

PROCEEDINGS REPRINT

SPIE—The International Society for Optical Engineering

Reprinted from

Medical Imaging 1996

Image Display

11–13 February 1996
Newport Beach, California

SPIE
P
PROCEEDINGS SERIES

Volume 2707

Box 10, Bellingham, Washington 98227 USA. Telephone 360/676-3290.

On Optimization of Orthonormal Wavelet Decomposition:
Implication of Data Accuracy, Feature Preservation, and Compression Effects

Shih-Chung B. Lo, Huai Li, Joseph Wang, Matthew T. Freedman, and Seong K. Mun

Center of Imaging Science and Information Systems, Radiology Department
Georgetown University Medical Center, Washington D.C. 20007
e-mail: lo@isis.imac.georgetown.edu

ABSTRACT

A neural network based framework has been developed to search for an optimal wavelet kernel that is most suitable for a specific image processing task. In this paper, we demonstrate that only the low-pass filter, h_u, is needed for orthonormal wavelet decomposition. A convolution neural network can be trained to obtain a wavelet that minimizes errors and maximizes compression efficiency for an image or a defined image pattern such as microcalcifications on mammograms. We have used this method to evaluate the performance of tap-4 orthonormal wavelets on mammograms, CTs, MRIs, and Lena image. We found that Daubechies' wavelet (or those wavelets possessing similar filtering characteristics) produces satisfactory compression efficiency with the smallest error using a global measure (e.g., mean-square-error). However, we found that Harr's wavelet produces the best results on sharp edges and low-noise smooth areas. We also found that a special wavelet, whose low-pass filter coefficients are (0.32252136, 0.85258927, 0.38458542, -0.14548269), can greatly preserve the microcalcification features such as signal-to-noise ratio during a course of compression. Several interesting wavelet filters (i.e., the *g* filters) were reviewed and explanations of the results are provided. We believe that this newly developed optimization method can be generalized to other image analysis applications where a wavelet decomposition is employed.

1. Introduction

In the field of transform coding, discrete cosine transform (DCT) based decomposition methods were developed extensively in 1970's and 1980's. Most of the techniques developed in this area are associated with block DCT[1-4]. However, several investigators indicated that the use of full-frame DCT[5-7] can produce high compression efficiency with high data fidelity and without blocky artifact. This method is particularly appropriate for high-resolution large-sized images. Recently, sub-band and wavelet transformations have been widely used in image compression research[8-10]. Unlike DCT, there exists many discrete wavelet transform (DWT) filters that can perform data decomposition. This paper provides a neural network approach to search for an optimal wavelet that minimizes quantization errors and at the same time produces the highest compression efficiency. This method can also be extended to evaluate various wavelets in preserving defined image features.

2. Algorithm Development

2.1. Construct a Neural Network using Wavelet Decomposition

The artificial neural network described in this paper is based on the convolution process which is used in the sub-band including wavelet decomposition. In fact, the wavelet-based neural network performs exactly the same as the conventional wavelet transform. Our approach is to use the training capability of the neural network to obtain the most suitable wavelet kernel for a specific signal processing task[11]. In this paper, our task is to minimize error and simultaneously achieve the highest compression efficiency during the course of compression and decompression processes. In order to

0-8194-2082-4/96/$6.00

match the sub-band decomposition, several characteristics of the neural network must be established: (a) no hidden but one output layer is used, (b) local connection through convolution process rather than fully connected nets is employed, and (c) the convolution process must be inversible (wavelet kernels are used in this paper). During compression and decompression processes, the inversible process is approximately conducted. The approximation is not due to the inverse transformation but because the inaccuracy of the quantized transform coefficients. Figure 1 shows the structure of the neural network using quantized transform coefficients as the targets.

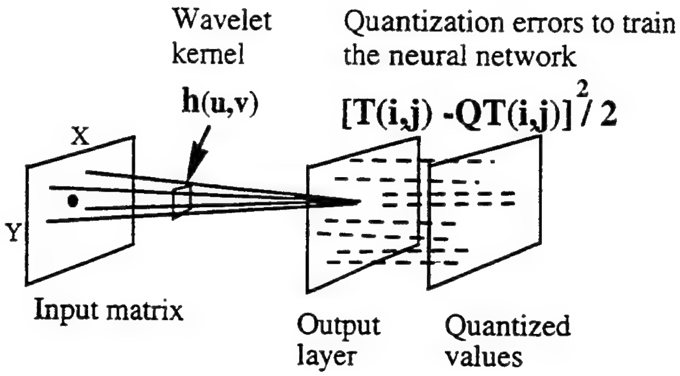


Figure 1. A neural network based on wavelet decomposition and trained by quantization errors. T(i,j) and QT(i,j) denote transform and quantized coefficients in high-frequency domains, respectively.

In fact, we should not consider only the issue regarding minimization of quantization errors. The minimization of entropy must also be taken into account for the optimization. We combine both issues by multiplying the mean-square-error function with an imposed entropy reduction function. The cost (error) function for training the neural network becomes

$$Ef(i,j) = Z(QT(i,j)) \times [T(i,j) - QT(i,j)]^2/2 \qquad \text{...(1)}$$

where $QT(i,j)$ is the quantized transform coefficient at pixel (i,j) and $Z(QT(i,j))$, which is the entropy reduction function for a set of quantization coefficients, is given below:

$$Z(QT(i,j)) = \begin{cases} 0 & for \quad QT(i,j) = 0 \\ 1 & for \quad |QT(i,j)| = 1 \\ F(n,q) & for \quad |QT(i,j)| = n \cdot \end{cases} \qquad \text{...(2)}$$

F(n,q), which is a ramp function, is a function of quantization factor, *q*, and is somewhat inversely proportional to the quantized integer, *n*. The value of the ramp function should always be smaller than 1.

The reason to design the entropy reduction function for a fixed quantizer, q, using eq. (2) is three-fold: (a) since most low value coefficients (-0.5q < T(i,,j) < 0.5q) are associated with noise when q is not a very large value, there is no need to emit error from the output node possessing quantized value 0 to train the neural net; (b) the more the low quantized values are, the lower the assemble entropy will be; and (c) the probability to turn a high quantized value into a low quantized value is very low, therefore errors backpropagated from high quantized values should be less emphasized as compared to low quantized values 1, 2, or so. When q is very small, the quantization error is in the range of global image noise. In this case, the neural network will rely on the guidance of Z function to search for a wavelet filter that produces more low transform values. The success of this cost (error) function design is depicted in our experiment shown in the Results Section.

Based on the neural network shown in Figure 1, we can train the convolution kernel. The specific training algorithm is given in Section 2.2. Unfortunately, the neural network suggested kernel may not be a wavelet kernel. Section 2.4 shows a method to conduct wavelet decomposition without using the high-pass filter. Hence, the low-pass filter is the only kernel to process the 4 channels for two-dimensional (2-D) wavelet decomposition. Section 2.5 provides algorithms that will modify the kernel to fulfill the requirements of wavelet kernel. Through this process, we can find a wavelet that produces the lowest quantization errors with the lowest entropy of the quantized transform coefficients.

2.2. Signal Propagation through Convolution Process and Methods for Training the Neural Network

The signal propagation from input layer to output layer involving convolution computation is given below:

$$T_c(i,j) = K_c(u,v) \otimes S(i,j) \qquad \text{...(3)}$$

where S(i,j) is the original image, subscript c denotes the channel number, and $K_c(u,v)$ is the convolution kernel for channel c. For the wavelet decomposition, the relationship between $K_c(u,v)$ and the wavelet filters (i.e., h and g filters) will be given in Sections 2.3 and 2.4.

Since we treat the wavelet transform as a locally connected neural network, the well-known backpropagation (BP) training method can be used to train the weights (kernel) in each epoch[12]. Note that a linear function instead of a typical sigmoid function for a conventional neural network system is used in this process. The updated kernel suggested by backpropagation in the neural network is given by

$$K_c(u,v)[t+1] = K_c(u,v)[t] + \eta \sum_{i,j} \delta(i,j) S(i-u, j-v) + \alpha \Delta K_c(u,v)[t] \qquad \text{...(4)}$$

where t is the iteration number during the training, α is the gain for the momentum term received in the previous learning loop, η is the gain for the current weight changes, and δ is the weight-update function which is given by

$$\delta(i,j) = \frac{\partial Ef}{\partial K_c(u,v)} . \qquad \text{...(5)}$$

2.3. Two-Dimensional Wavelet Decomposition

Following Mallat's 2-D wavelet analysis[9], the two-dimensional scaling function is composed of two one-dimensional scaling functions in both directions:

$$\phi(x,y) = \phi(x)\phi(y) \qquad \text{...(6)}$$

where $\phi(x)$ is a scaling function. The associated two-dimensional wavelets are defined as

$$\psi^H(x,y) = \phi(x)\psi(y) \qquad \text{...(7)}$$

$$\psi^V(x,y) = \psi(x)\phi(y) \qquad \text{...(8)}$$

$$\psi^D(x,y) = \psi(x)\psi(y) \qquad \text{...(9)}$$

where $\psi(x)$ is the 1-D wavelet corresponding to the 1-D scaling function. Using the sub-band coding algorithm, the wavelet transform (2-D DWT) of a matrix has four parts:

$$W_{LL}(f(x,y)) = \sum_{u,v}\left[(f(x,y)h(u-2x,0))h(0,v-2y)\right] = \sum_{u,v}\left[f(x,y)h_{LL}(u-2x,v-2y)\right] \quad ...(10)$$

$$W_{LH}(f(x,y)) = \sum_{u,v}\left[(f(x,y)h(u-2x,0))g(0,v-2y)\right] = \sum_{u,v}\left[f(x,y)h_{LH}(u-2x,v-2y)\right] \quad ...(11)$$

$$W_{HL}(f(x,y)) = \sum_{u,v}\left[(f(x,y)g(u-2x,0))h(0,v-2y)\right] = \sum_{u,v}\left[f(x,y)h_{HL}(u-2x,v-2y)\right] \quad ...(12)$$

$$W_{HH}(f(x,y)) = \sum_{u,v}\left[(f(x,y)g(u-2x,0))g(0,v-2y)\right] = \sum_{u,v}\left[f(x,y)h_{HH}(u-2x,v-2y)\right] \quad ...(13)$$

where *h* and *g* functions are the low and high pass filters of the sub-band decomposition with condition $g(u) = (-1)^u h(1-u)$. The low pass filter, *h*, also must satisfy three criteria to construct the orthonormal basis of compactly supported wavelets[5,6]: (Note that we also use g_u and h_u to replace *g(u)* and *h(u)*, repectively, for simplicity in this paper.)

(a) $$\left[\sum_u h_{2u}\right] - \sqrt{2}/2 = \left[\sum_u h_{2u+1}\right] - \sqrt{2}/2 = 0; \quad ...(14)$$

(b) should be orthonormal; this means that

$$\left[\sum_u h_u \times h_{u+2n}\right] - \delta_{u,u+2n} = 0 \quad ...(15)$$

where $\delta_{i,j}$ is Dirac delta function and *n* is an integer; and

(c) have a high degree of regularity.

From the compression perspectives, the above constraints are very limited. For a lossless compression, those filers performing perfect reconstruction are illegible. However, we would like to focus our view on using wavelet transform in this paper.

The 2-D filters at the second forms of eqs. (10-13) are the vector products of *h* and/or *g* filters. The relationship between high pass and low pass filters make the unification of the four sets of decomposition possible as shown in section 2.4.

According to the wavelet theory, it is known that given a set of *h*, one can calculate the Fourier transform of the scaling and wavelet functions as follows:

$$\Phi(w) = H_0(e^{iw/2})\Phi(w/2) \quad ...(16)$$

$$\Psi(w) = H_1(e^{iw/2})\Phi(w/2) \quad ...(17)$$

where H_0 and H_1 are Fourier transforms of *h* and *g* filters, respectively. Hence, both the scaling and wavelet functions can be obtained through infinite recursion by using eqs. (16) and (17), respectively.

2.4. Unification of the Four Channels Decomposition in 2-D DWT

Using Eq. (11) as an example to rewrite the decomposition equation by replacing the *g* with the *h* filter, we have:

$$W_{LH}(f(x,y)) = \sum_{u,v}\left[(f(x,y)h(u-2x,0))(-1)^v h(0,2y+1-v)\right] \quad ...(18)$$

or

$$\begin{aligned} W_{LH}(f(x,y)) &= \sum_{u,v}\left[(((-1)^v f(x,-y))h(u-2x,0))h(0,v-2y)\right] \\ &= \sum_{u,v}\left[(((-1)^v f(x,-y))h_{LL}(u-2x,v-2y)\right] = \sum_{u,v}\left[f_{LH}(x,y)h_{LL}(u-2x,v-2y)\right]. \end{aligned} \quad ...(19)$$

Converting Eq. (14) to use the 2-D low pass filter as the kernel is a matter of changing the orientation from y- to x-direction (or combining both directions for Eq. (15)). These conversions also indicate that one can use a single 2-D filter to compute the four quadrants of the 2-D wavelet transform by flipping the matrix position in x- and/or y-direction(s) and alternating the sign of the flipped matrix corresponding to the direction(s).

The alternated sign of the source vector makes the convolution operation unconventional. A precalculation method, that involves a cross product of two vectors, can be employed: flipping data sequence of an image is the first vector and the second vector is fixed and composed of +1 and -1. An example of 1-D precalculation steps for tap-6 kernel prior to the convolution operation is given below:

Original data sequence:	$a_1, a_2, a_3, a_4, a_5, a_6$
Flipped data sequence:	$a_6, a_5, a_4, a_3, a_2, a_1$
Resultant data sequence:	$a_6, -a_5, a_4, -a_3, a_2, -a_1$

In the case of 2-D, three matrices associated with horizontal, vertical, and diagonal decomposition for the second matrix in precalculation are given below in Figure 2. With this precalculation (or cross product of two matrices), only the low-pass filter $h_u h_v$ (h_u in 1-D) is needed for the final wavelet transform operation.

$$\begin{bmatrix} + & + & + & + & + & + \\ - & - & - & - & - & - \\ + & + & + & + & + & + \\ - & - & - & - & - & - \\ + & + & + & + & + & + \\ - & - & - & - & - & - \end{bmatrix} \qquad \begin{bmatrix} + & - & + & - & + & - \\ + & - & + & - & + & - \\ + & - & + & - & + & - \\ + & - & + & - & + & - \\ + & - & + & - & + & - \\ + & - & + & - & + & - \end{bmatrix} \qquad \begin{bmatrix} + & - & + & - & + & - \\ - & + & - & + & - & + \\ + & - & + & - & + & - \\ - & + & - & + & - & + \\ + & - & + & - & + & - \\ - & + & - & + & - & + \end{bmatrix}$$

Vertical operator *Horizontal operator* *Diagonal operator*

Figure 2. Three matrices used for the cross product precalculation.

Nevertheless the resultant matrix of this precalculation (or cross product of two matrices) must be held in the computer memory to facilitate the computation for forward convolution and the corresponding backpropagation. After precalculation, the size of the intermediate images is $(k/2 \times k/2)$ times the original image size. The factor of $1/2 \times 1/2$ is due to the 1/2 down sampling two-dimensionally in a conventional forward wavelet transform. The largest three blocks shown in Figure 3 are the intermediate images $S_O(xk/2, yk/2)$.

One of the original criteria regarding the so-called "high degree of regularity" was not enforced in the algorithm. The orthonormality of the h_u filter may not be self-sustained with each updated version. However, some small modification is possible to make the final version of h_u orthonormal, if the conditions of being a wavelet filter set are to be fully met. Based on each precalculated image $S_O(xk/2, yk/2)$ described earlier, Eq. (4) can be rewritten for updating 2-D wavelet kernel

$$K(u,v)[t+1] = K(u,v)[t] + \eta \sum_{i,j} \delta(i,j) S_0(xk/2-u, yk/2-v) + \alpha \Delta K(u,v)[t] \qquad ...(20)$$

where index i = 0,1,...$(k-1)^2$ corresponds to the sub-image of S_o matched to the kernel size. Eq. (20) represents the updated kernel suggested by the BP, these values require a conversion to a new wavelet kernel $h'_u h'_v$. Assuming the wavelet filter is a 2-D vector (i.e., $h_u h_v = h_v h_u = h_{LL}$, where $u\&v = 0,1,2,$ ... k-1), then only k free parameters ought to be trained for a wavelet transform. A solution to satisfy the wavelet constraints and to make $h'_u h'_v$ approximately equal to $K'(u,v)$ is given in section 2.5.

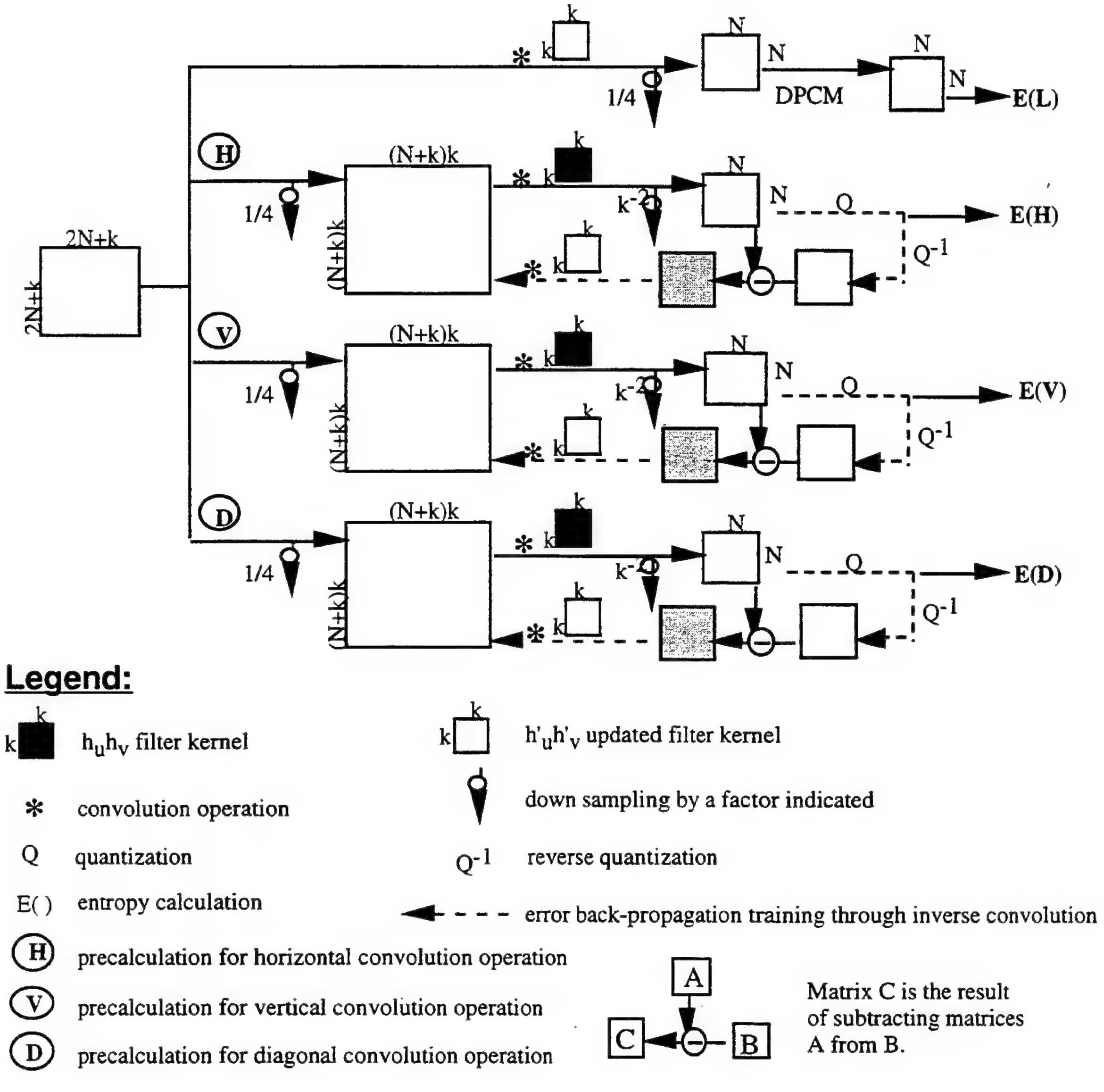

2N+k
2N+k
H
V
D
1/4
(N+k)k
(N+k)k
k
N
DPCM
E(L)
k-2
Q
Q-1
E(H)
E(V)
E(D)
Legend:
hu hv filter kernel
h'u h'v updated filter kernel
convolution operation
down sampling by a factor indicated
quantization
reverse quantization
entropy calculation
error back-propagation training through inverse convolution
precalculation for horizontal convolution operation
precalculation for vertical convolution operation
precalculation for diagonal convolution operation
A
B
C
Matrix C is the result of subtracting matrices A from B.

Figure 3. A proposed training scheme based on a grouped (kernel) backpropagation neural network to obtain an optimal orthonormal kernel for image compression.

2.5. Converting Neural Network Suggested Kernel to Fulfill Requirements of a Wavelet Filter

As indicated in Eq. (20), the updated weights, $K(u,v)[t+1]$ or $K'(u,v)$ of the kernel suggested by the BP at $t+1$ training iteration are independent. One must realize that each epoch in the neural network training is only a suggestion or approximation that the changes of weights may produce a lower value for the defined error function, Ef. To properly use this suggestion for making a new wavelet kernel, let's assume that there exists a set of h'_u so that the updated 2-D version of the wavelet filter is very close to $K'(u,v)$. A function based on the square difference is used in the derivation

$$f(h'_u) = \sum_{u,v} (h'_u h'_v - K(u,v))^2. \qquad ...(21)$$

Here we intend to minimize the function, *f*, subject to the constraints equations. Lagrangian multiplier method can be employed to solve this problem by combining *f* and constraint equations:

$$df(h'_u) + \sum_p \lambda_p dC_p(h'_u) = 0 \qquad ...(22)$$

where d represents the differentiation operation of a function and λ_p is the Lagrangian multiplier for the corresponding constraint equation, $C_p(h'_u) = 0$, referred to eqs. (14) and (15). Using this approach we can obtain a set of h'_u while *f* is also minimized.

3. Materials and Experimental Methods

A database consisting of 45 mammograms was used to conduct the study. Of these mammograms, 38 contain biopsy proven clustered microcalcifications. A total of 220 microcalcifications are embedded in 41 clusters. All 45 mammograms were digitized by a LumyScan (model 150) film digitizer with spot size of 100µm. Each patch of 32×32 pixels (i.e., an area of 3.2×3.2 mm^2) with its center at the peak value was isolated for the study of quantization impact on microcalcifications. The process of searching optimal wavelet kernels for original mammograms and microcalcification patches were conducted. Each image was decomposed by 3-level wavelet transform. Quantization values were q, q/2, and q/4 for decomposition of high frequency coefficients on levels 1, 2, and 3, respectively. For each training epoch, the mean-square-error (MSE) and %zeros (i.e., number of zeros / total number of pixels) were computed. Since %zeros generally contributes the most important factor to gain a compression, it can be used as a coarse index for the evaluation of compression efficiency for each epoch.

In order to demonstrate each wavelet performance, we sorted the first coefficient h_0 of the low-pass filter associated with the mother scale function as the horizontal scale because the training epoch does not represent the wavelet being used as shown in Figures 6 and 7. All h_0 values are greater than -0.1464466094 and smaller than 0.85255533905. The corresponding h_1 values are greater than 0.35355339 and smaller than 0.85255533905. Those h_1 values, which are greater than -0.1464466094 and smaller than 0.35355339, have corresponding conjugate values in the former set and can be ignored.

Compression ratios were calculated only when the neural network search had been successful. The spatial and temporal correlation of quantized coefficients were taken into account but might not be optimized. Specifically, we arranged quantized coefficients from one pixel of the highest level to the corresponding 4 pixels on the second highest level to 16 pixels on the lowest level and then went back to the next pixel of the highest level and so on. This rearranged data sequence is more correlated in a spatial-temporal sense[13] and can be encoded effectively by Lempel-Ziv coding[14].

We have also performed the same study for the isolated 220 microcalcification patches. The 2-D profiles of microcalcifcations and their nearby areas (i.e., the areas that are not included in the microcalcification profile but within the isolated block 32×32 pixels) were evaluated separately during the course of the neural network search. In addition, features of the microcalcifications were computed to observe their changes. These features of microcalcification are:

(a) the peak value, P;

(b) the contrast, C = P-b;
where b is the average background value which is the immediate boundary of the microcalcification profile;

(c) the signal-to-noise-ratio, SNR = C/SD$_b$;
where SD$_b$ stands for the standard deviation of the background; and

(d) the area occupied by the 2-D microcalcification profile, A.

4. Results

In the neural network training, the MSE is not the only factor to be concerned; the entropy reduction function is another factor that drives the neural network to perform a search. In the first neural network experiment, we found that the MSE changes very small with a low quantization factor (q=16). The neural network movement in searching for the next wavelet kernel was random and no minimum of MSE could be found in the mammogram study. However, the %zeros changed which led the neural network to converge at the maximum value of %zeros. When a larger quantization factor (q=64) was used, the MSE seems to function in training the neural network. Figures 4 and 5 show the curves of MSEs and %zeros against the sorted h_0 values. In both figures, Daubechies' ($h_0 = 0.48296291$) and its nearby wavelets perform the highest %zeros implying the largest compression ratio.

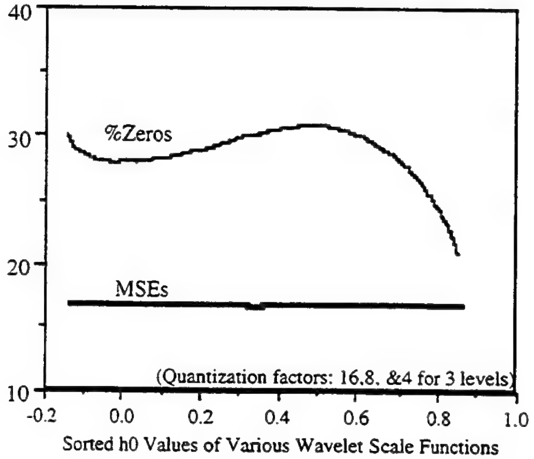


Figure 4. Decomposition Performance of Wavelets on Mammograms (q=16).

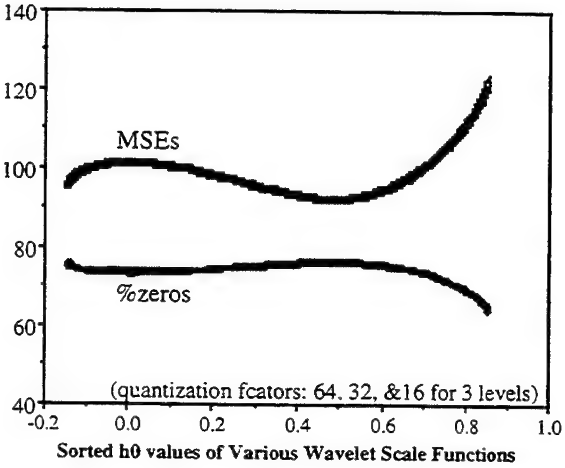


Figure 5. Decomposition Performance of Wavelets on Mammograms (q=64).

In the microcalcification study, we found that %zeros does not change much until $h_0 > 0.6$. Figure 6 shows the original learning steps which drive MSEs into lower values using the proposed neural network training mechanism. Figure 7, which is a sorted version of Figure 6 (same sorting processes were applied to all figures in this section), shows that Daubechies' wavelets perform the lowest MSEs. More specifically, microcalcification profiles suffered higher MSEs than their background areas as indicated in Figure 8.

These results were altered when a very large quantization factor was used. In Figure 9, all the microcalcification patches were rounded-off to 8-bit prior to the study which assumed digitized mammograms containing about 4-bit of noise[15]. Although the largest quantization factor was 16 for 8-bit mammograms, the effective quantization factor was equivalent to ≈256 in 12-bit mammograms. Figure 9 shows that Harr's wavelet ($h_0 = 0.0$) performs a high and the lowest MSEs for 2-D microcalcification profiles and their background, respectively. However, Daubechies' wavelets perform in an opposite way. This is probably because Harr's wavelet can produce a lower entropy for low-noise smooth areas.

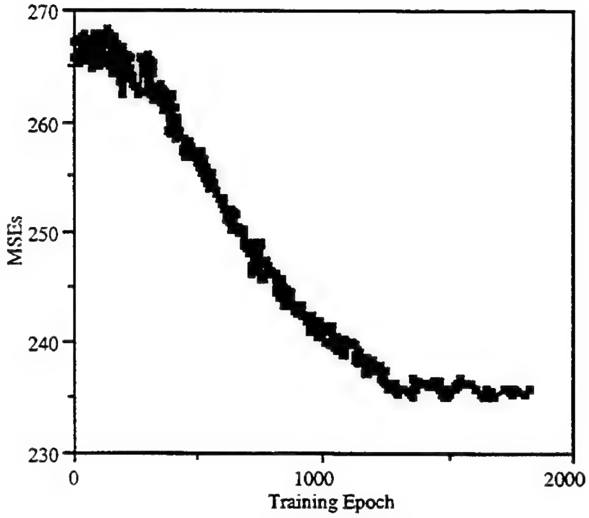


Figure 6. MSEs were Decreased During the Training of the Neural Network on 220 Microcalcifications (q=64).

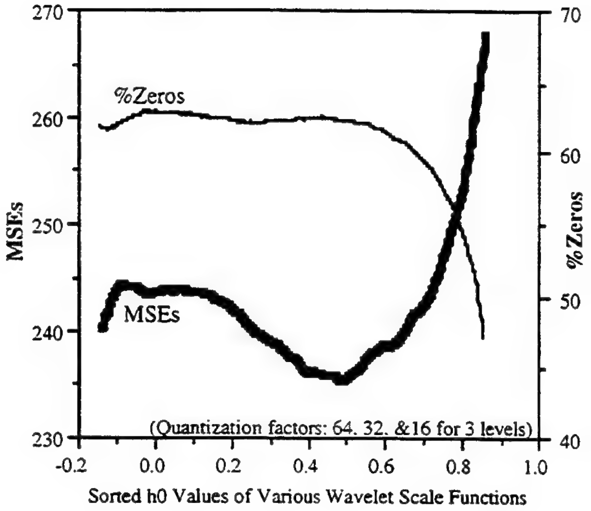


Figure 7. Decomposition Performance of Wavelets on 220 Microcalcifications (q=64).

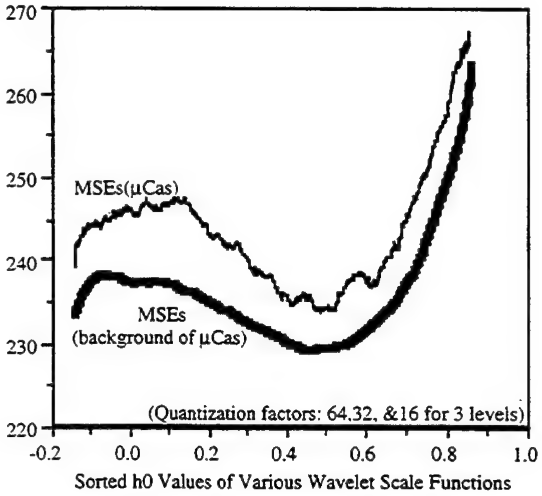


Figure 8. Decomposition Performance of Wavelets on 220 Microcalcification Profiles and Background (q=64).

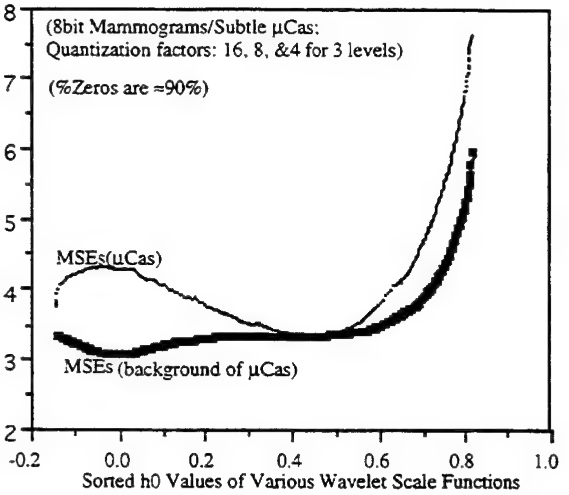


Figure 9. Decomposition Performance of Wavelets on 220 Microcalcification Profiles and Background (8-bit, q=16).

The results of the microcalcification evaluation study based on quantized wavelet coefficients are shown in Figures 10-13. In fact, the evaluation was performed with an identical experimental condition as that in Figure 9. However, microcalcification features were measured instead of MSEs and %zeros. Note that % number decrease in peak values, contrast, and SNR were shown in negative values. In other words, the lower the % number decrease value is, the more microcalcifications involving negative changes. The figure of merit (FOM) for each measure was a composed value given by

$$\text{FOM} = (\ \%\text{No. decrease} \times \%\ \text{decrease} + \%\text{No. increase} \times \%\ \text{increase}) \times 100. \qquad ...(23)$$

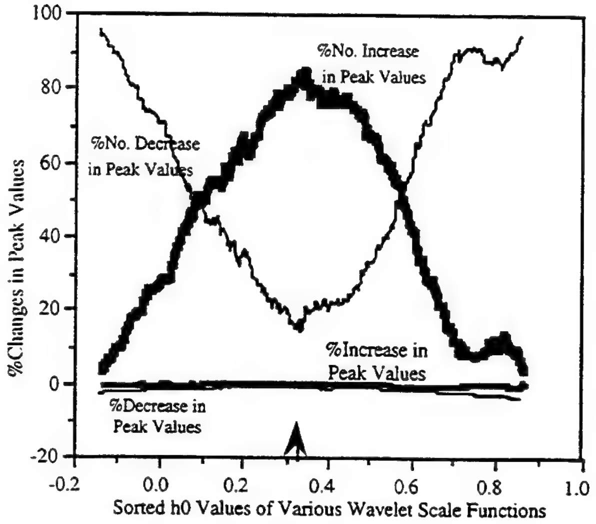


Figure 10. Peak Values Changes Due to Quantization Effects on Wavelet Domain for Microcalcifications.

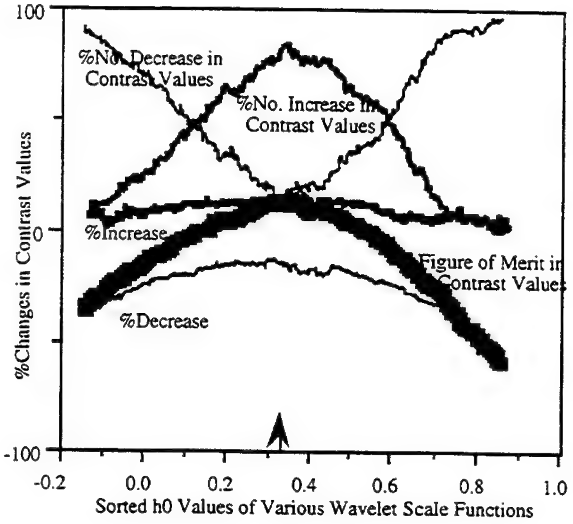


Figure 11. Contrast Changes Due to Quantization Effects on Wavelet Domain for Microcalcifications.

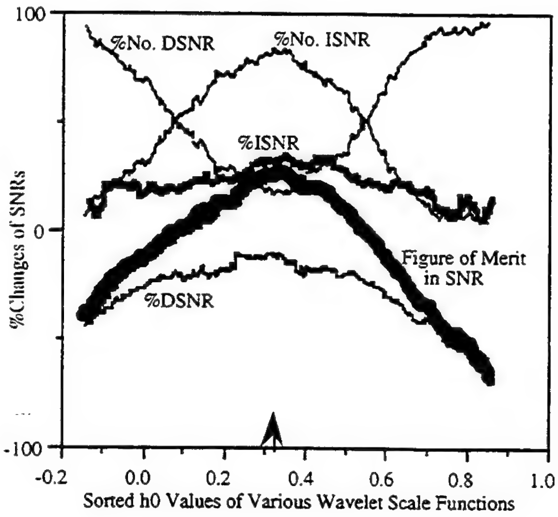


Figure 12. SNR Changes Due to Quantization Effects on on Wavelet Domain for Microcalcifications.

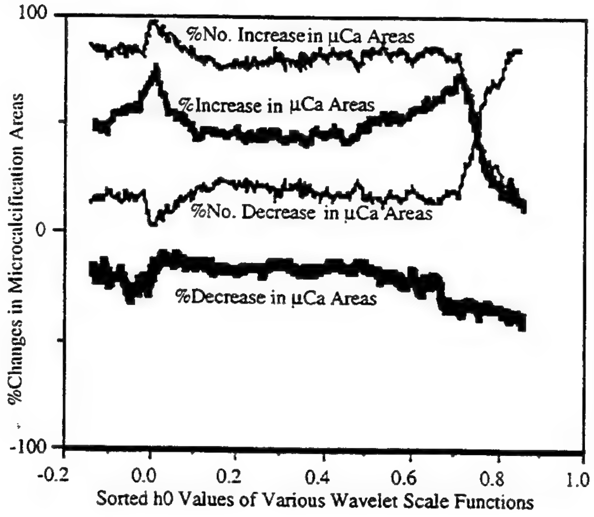


Figure 13. Areas of Microcalcification Profile Changes Due to Quantization Effects on Wavelet Domain.

As indicated in Figure 10, the peak values were changed very little. However, % number increases in peak values, contrast values, and SNRs of microcalcifications had about the same distribution in Figures 10, 11, and 12. The highest FOMs in all three measures were at the wavelet with the low-pass filter coefficients: (0.32252136, 0.85258927, 0.38458542, -0.14548269) which is marked with an arrow sign in the Figures. Figure 13 shows minor %area changes of microcalcification profiles from 0.2 to 0.6 of h_0 values. These effects were not observed when a low quantization factor was used.

5. Discussion

After observing these results, one may still be confused about what was going on. The authors would like to provide some explanations in the following discussion. Let's start with graphics of the low-pass *h* and the high-pass *g* filters for several interesting wavelets mentioned in the Results Section. Figures 14 and 15 show the h and g filters, respectively. Note that the X wavelet is the same wavelet marked on the horizontal axis in Figures 10, 11, and 12. One should pay more attention to the graphics of the *g* filters, since they produce high frequency coefficients for quantization. We can deem that *g* filter essentially performs calculation involving the positive weight multiplied by the center pixel value plus the adjacent pixel values on the two sides multiplied by the negative weights of the *g* filter. Daubechies' wavelet has quite balanced negative terms at the two sides of the positive weight and the sum of negative weights is negatively equal to the positive weight. The latter is a constraint in all wavelet filters anyway. In addition, the absolute value of $g1(=-h2)$ or $g2\ (=h1)$ should be reasonably large, which would maintain the low-pass and the high-pass characteristics for *h* and *g* filters, respectively. In fact, those wavelets near Daubechies' wavelet including the one (X wavelet) with the highest performance in microcalcification features possess this property. From the signal processing point of view, these balanced weights in a filter are very important characteristics to create low entropy values for general textures. We suspect that this property may have something to do with the so called "high regularity" in the wavelet theory.

In short, we found that the main reason that a wavelet filter can produce a low entropy for a set of data is because the weight sum of the *g* filter is zero. For a general data sequence, the *g* filter can perform even better when

(a) the absolute value of $g1(=-h2)$ or $g2(=h1)$ is much larger than that of other weights.
(b) the opposite signed weights are evenly distributed at the two sides of $g1$ or $g2$.

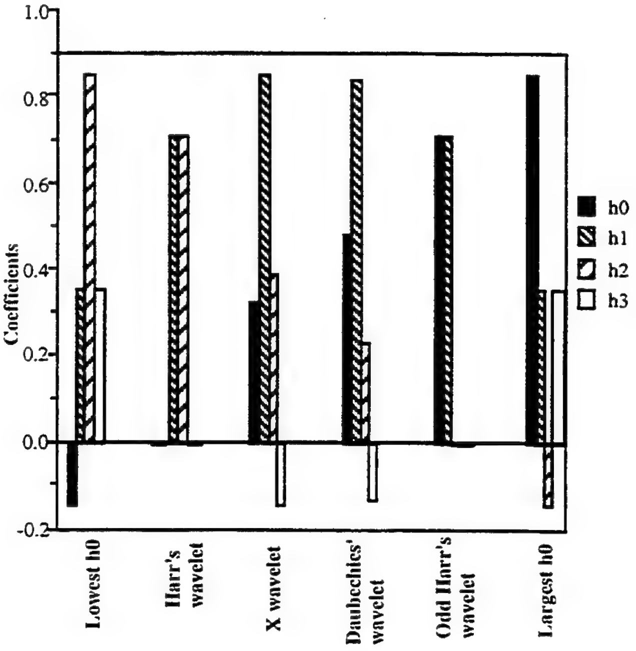


Figure 14. Low-pass Filters of Several Interesting Wavelets.

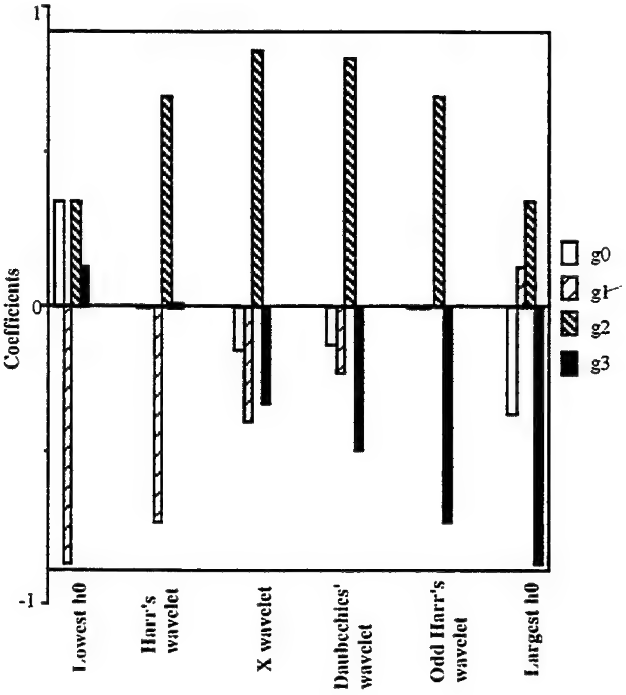


Figure 15. High-pass Filters of the Same Wavelets.

For low-noise smooth signals, Harr's wavelet may slightly outperform the others. For sharp edges, Harr's wavelet would greatly outperform the others, as depicted in Figure 16 where only bones as well as edges between bones and soft tissues isolated on computed tomographic (CT) images were the subjects for the evaluation.

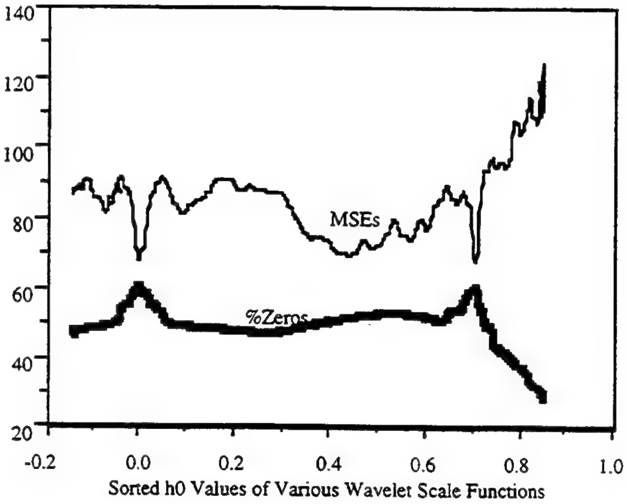


Figure 16. Decomposition Performance of Wavelets on CT Head Bones and Bone Edges (q=64).

We still do not quite understand why the wavelet possessing low-pass filter (0.32252136, 0.85258927, 0.38458542, -0.14548269) resulted in the highest feature preservation. However, Figure 9 has provided clues as to where MSEs of 2-D microcalcification profiles and background gradually merge from Harr's to Daubechies' wavelets. Since contrast and SNR values are computed using the peak and background values of the microcalcifications, the optimization of these measures should occur somewhere between Harr's and Daubechies' wavelets.

In the field of compression, it is known that the higher the compression ratio is, the higher the error that will be generated in the decompressed image. However, through these studies we discovered a new phenomenon associated with these two main quantitative measures in compression. We found that higher compression coincided with less error in all the studies (see Figures 4, 5, 7, & 16) using a fixed quantizer. This may be because high compression is associated with low entropy, which means that the data contains more low values and less variation between the originally transformed and quantized coefficients. This phenomenon happens only when the quantization factor is fixed. We would like to call for the reader's attention to the link between this phenomenon and the designed error function that comprises MSE and entropy reduction terms for training the convolution neural network. With this concurrent trend (i.e., less error is associated with low entropy using a fixed quantizer), the neural network can be effectively trained. Otherwise they would have functioned as competing factors and would have made the training of the neural network difficult.

Although we have shown the general framework of a wavelet filter search using a neural network training method, only tap-4 wavelet spectra were employed in our experiment. The above research findings seem able to be generalized for high order wavelets because the *g* filter is the key operator for the wavelet decomposition. The distribution of weights for high order wavelets should be maintained as discussed above in order to obtain a low entropy. We will continue to investigate the performance of bi-orthonormal wavelets where an odd number of weights are used. We predict that high performance wavelets in compression and data accuracy should possess balanced distribution of weights in the *g* filter of bi-orthonormal wavelets.

In our previous papers, we indicated that wavelet (both orthonormal and bi-orthonormal) decomposition might be appropriate for low resolution small images such as the Lena image, CTs and MRIs. For high resolution large images such as digitized chest radiographs and mammograms, we found that the full-frame DCT performed with the highest compression efficiency[16]. This is because the DCT can pack highly correlated image information in a small frequency area. The DWT, however, requires many levels in decomposition to achieve a high compression ratio. The data inaccuracy would propagate from high level wavelet domains to low level and to the reconstructed image.

6. Conclusions

A neural network based method has been developed to search for optimal wavelet kernels which can produce the most favorable set of transform coefficients to preserve data accuracy and/or defined image features during the compression. In this paper, our technical achievements are: (a) development of a unified method to facilitate multichannel wavelet decomposition; (b) designing a cost (error) function consisting of MSE and imposed entropy reduction function for training the convolution neural network; and (c) converting neural network suggested kernel into a filter constrained by the wavelet requirements.

In all medical image modalities we have tested so far (including mammography, CT, MRI), Daubechies' wavelet (or its nearby wavelets) generally performs better (in most cases slightly better) than other wavelts for image compression using a global measure. With a large quantization factor, Harr's wavelet produces the lowest and highest MSEs for the background and microcalcification profile areas, respectively. However, Daubechies' wavelet produces an opposite result. In addition, we found that the wavelet associated with a low-pass filter, (0.32252136, 0.85258927, 0.38458542, -0.14548269), possesses the highest feature preservation capability in microcalcification peak, contrast, and SNR. Through this study, we also found that only Harr's wavelet sometimes produced a dramatic result, usually optimization occurs on a band of wavelets not at a single wavelet.

We, therefore, conclude that Daubechies' wavelet (and its nearby wavelets) is generally applicable for image compression. However, Harr's wavelet is suitable for low-noise smooth areas and sharp edges. For a specific image pattern such as microcalcifications on mammograms, one might find a wavelet filter can most preserve the features.

By reviewing the *g* filters of various wavelets, we found those optimal wavelets for general image texture have something in common. They possess balanced negative terms at the two sides of the positive weight and the absolute value of *g1* or *g2* is much larger than that of the other weights.

7. Acknowledgments

This work was supported by a NIH/NCI Grant No. 5R01CA59763 and US Army Research Grant No. DAMD17-93-J-3007. The content of this paper does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. The authors are grateful to Ms. Susan Kirby for her editorial assistance.

8. References

1. Pratt,WK: Digital Image Processing, John Wiley & Sons, New York, 1978.
2. Jain AK, "Image Data Compression: a review," *Proc. IEEE*, vol. 69, 1981, pp. 349-389.
3. Rosenfeld A and Kak AC, "Digital Picture Processing," Acdemic Press, 1982.
4. "Initial Draft for Adaptive Discrete Cosine Transform Technique for Still Picture Data Compression Standard," ISO/IEC JTC1/SC2/WG8 N800, JPEG Technical Specification, Revision 8, Aug. 1990.
5. Lo SC and Huang HK, "Compression of Radiological Images with Matrix Sizes 512, 1024, and

2048," Radiology 1986, pp. 519-525.

6. Huang HK, Lo SC, Ho BK, and Lou S: "Radiological Image Compression using Error-Free an Irreversible Compression and Reconstruction in 2-Dimensional Direct Cosine Transform Coding Techniques. J. Optical Society of America. May 1987, pp. 984-992.
7. Lo SC, Krasner BH, Mun SK, and Horii SC, "The Full-Frame Entropy Encoding for Radiological Image Compression," SPIE Proc. Medical Imaging V, Vol. 1444, 1991. pp. 265-277.
8. Daubechies I, "Orthonormal Based of Compactly Supported Wavelets", Comm. on Pure and Appl. Math., Vol. XLI, 1988, pp. 909-996.
9. Mallat S, "A Theory For Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Trans. Pat. Anal. Mach. Intel., Vol. 11 No. 7, 1989, pp. 674-693.
10. Antonini M, Barlaud M, Mathieu P, Daubechies I, "Image Coding Using Wavelet Transform," IEEE Trans. Image Proc., vol. 1 No. 2, 1992, pp. 205 - 220.
11. Lo SC, Li H, Lin JS, Hasegawa A, Wu YC, Freedman MT, and Mun SK, "Artificial Convolution Neural Network with Wavelet Kernel for Disease Pattern Recognition," SPIE proceedings, Medical Imaging, 1995, Vol, 2434, pp. 579-588.
12. Rumelhart DE, Hinton GE, & Williams RJ (1986). Learning internal representation by error propagation. In D.E. Rumelhart & J.L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing,* **1** (pp. 318-362). Cambridge, MA: MIT Press.
13. Xiong Z, Ramchandran K, and Orchard MT, and Asai K, "Wavelet Packets-Based Image Coding using Joint Space-frequency Quantization, In Proc. IEEE Int. Conf. on Image Proc. Austin, TX, 1994, pp. 324-328.
14 Ziv J and Lempel A, "A Universal Algorithm for Sequential Data Compression," IEEE Trans. on Info. Theory, Vol. IT-23, No. 3, May 1977, pp. 337-343.
15. Lo SC, Krasner BH, and Mun SK, "Noise Impact on Error-Free Image Compression," IEEE, Trans. Medical Imaging, Vol. 9, No. 2, June 1990, pp. 202-206.
16. Lo SC, Li H, Krasner BH, Freedman MT, and Mun SK, "Large-Frame Compression using DCT and Wavelet Transform Techniques," SPIE proceedings, Medical Imaging, 1995, Vol 2431, pp. 195-202.

Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue

Berkman Sahiner, Heang-Ping Chan, Datong Wei, Nicholas Petrick, Mark A. Helvie,
Dorit D. Adler, and Mitchell M. Goodsitt
The University of Michigan, Department of Radiology, Ann Arbor, Michigan 48109-0030

(Received 5 July 1995; resubmitted 14 May 1996; accepted for publication 3 July 1996)

We investigated a new approach to feature selection, and demonstrated its application in the task of differentiating regions of interest (ROIs) on mammograms as either mass or normal tissue. The classifier included a genetic algorithm (GA) for image feature selection, and a linear discriminant classifier or a backpropagation neural network (BPN) for formulation of the classifier outputs. The GA-based feature selection was guided by higher probabilities of survival for fitter combinations of features, where the fitness measure was the area A_z under the receiver operating characteristic (ROC) curve. We studied the effect of different GA parameters on classification accuracy, and compared the results to those obtained with stepwise feature selection. The data set used in this study consisted of 168 ROIs containing biopsy-proven masses and 504 ROIs containing normal tissue. From each ROI, a total of 587 features were extracted, of which 572 were texture features and 15 were morphological features. The GA was trained and tested with several different partitionings of the ROIs into training and testing sets. With the best combination of the GA parameters, the average test A_z value using a linear discriminant classifier reached 0.90, as compared to 0.89 for stepwise feature selection. Test A_z values with a BPN classifier and a more limited feature pool were 0.90 with GA-based feature selection, and 0.89 for stepwise feature selection. The use of a GA in tailoring classifiers with specific design characteristics was also discussed. This study indicates that a GA can provide versatility in the design of linear or nonlinear classifiers without a trade-off in the effectiveness of the selected features.

Key words: mammography, computer-aided diagnosis, genetic algorithms, feature selection

I. INTRODUCTION

Computer-aided diagnosis (CAD) for detection and classification of breast abnormalities on mammograms is an active area of research.[1] Clinical studies have shown that 10% to 30% of breast cancers visible on mammograms in retrospective studies were initially missed by radiologists,[2,3] and that only 15% to 30% of the patients who have undergone biopsy due to a suspicious finding on mammograms are found to have breast cancer.[4,5] CAD methods have the potential of reducing the false-negative rate while improving the positive predictive values of the mammographic abnormalities.

Masses are important indicators of malignancy on mammograms. In recent years, considerable effort has been devoted to the development of computerized methods for detection and classification of masses.[6–12] In all of these investigations, the detection or classification task relies on the use of features extracted from the digitized mammograms. The extracted features represent properties of pixels (or groups of pixels) which contain characteristic information of the masses. In this paper, we report our development of a computerized method for classification of regions of interest (ROIs) on mammograms as either masses or normal tissue, with particular emphasis on a genetic algorithm for feature selection.

Feature selection is a very important step in classification,[8,10,11,13–16] because the inclusion of inappropriate features often adversely affects classifier performance, especially when the training set is not sufficiently large. The methods employed for feature selection vary. In some approaches,[7,9] very few features were used, and the process of feature selection was not clearly described. It is reasonable to assume that the features were selected on the basis of some prior knowledge from clinical experience. Wu *et al.*[13] selected 14 features from a total of 43 for classification of malignant and benign masses, and observed an improvement in classification accuracy when the reduced feature space was used instead of the entire feature space. The criterion for selection was the difference of the average values of individual features between the two classes. Goldberg *et al.*[14] first selected five features from a total of 26 based on the ability of the individual features to discriminate between malignant and benign masses. Subsequently, based on their pairwise discriminatory ability, three final features were selected from the remaining five features. In the study by Chitre *et al.*,[15] the criterion for texture feature selection was the combination of a classification error and a clustering technique using individual features independently. In our previous studies, we employed a stepwise feature selection procedure in linear discriminant analysis (LDA),[10,11] in which a feature is included or excluded at each step based on a chosen statistical criterion. The LDA takes into account the correlation between the features and the joint probability distribution of the feature vectors in the multidimensional feature space.

Many feature selection methods have been explored in CAD. However, the best method which can provide the highest accuracy for a given application is still in question. This is partly because feature selection is theoretically a difficult problem.[17] It is well known, for example, that the two independent features that yield the highest classification accuracy in a feature set may not constitute the best pair of features together.[18] In the training process in CAD, the classifier can be designed so that the probability of training error will not increase when the number of selected features increases. However, when both training and testing are desired, the problem becomes more complicated due to overfitting. Test results can deteriorate when the number of selected features increases,[13] especially when the number of training cases is small. It is imperative to select a smaller subset of features to overcome the so-called "curse of dimensionality"[19,20] (decrease in classification accuracy of the test set with an increasing number of features) if the ratio of the number of training cases to the number of available features is not sufficiently large. Several recipes for feature selection are mentioned in the literature,[19,21] but none of these, except for an exhaustive search procedure, is optimal.

Genetic algorithms (GAs), first introduced by Holland in the early seventies,[22] are becoming increasingly popular in solving optimization and machine learning problems.[23,24] The fundamental principle underlying GAs is based on natural selection. To solve an optimization task, a GA maintains a population of bit strings, which are referred to as chromosomes. Each chromosome corresponds to a possible solution of the problem. In each generation of the GA, the population is probabilistically modified, generating new chromosomes which may have a better chance of solving the optimization problem. GAs have been applied to complex optimization problems such as the control of a gas-pipeline system,[25] design of jet engine turbines,[26] training of a backpropagation neural network,[27] feature selection for an artificial neural network,[28] and automated detection of lung nodules.[29] GAs usually yield nonoptimal, but near-optimal solutions. They are thus well-suited for feature selection problems in large feature spaces, where the optimal solution is practically impossible to compute, and a near-optimal solution is the best alternative.

In this paper, we studied the ability of a GA to select features from a large feature space. Our goal was to introduce a more effective and versatile feature selection mechanism. The effectiveness and the versatility of the GA was demonstrated by its application to the problem of classification of masses and normal tissue on mammograms. The feature space included local and global multiresolution texture features[30] as well as morphological features.[31] The rest of the paper is organized as follows. In the next section, we briefly discuss important components of a GA. In Sec. III, we describe our image database, background correction method, extraction of texture and morphological features, and the GA implementation for feature selection. In Sec. IV, we evaluate the dependence of the classification results on different GA parameters. Section V contains a discussion of these results. Finally, Sec. VI concludes the investigation and provides a scope for further research.

II. GENETIC ALGORITHMS

In natural evolution, the basic problem of each population is to find beneficial adaptations to a complex environment. The characteristics that each individual has gained or inherited are carried in its chromosomes and each individual reproduces more or less in proportion to its fitness within the environment. Crossover and mutation provide the possibility of evolution toward better-fit individuals.

Genetic algorithms[22–24] apply the principles of natural selection to machine learning. To solve an optimization problem, a GA requires five components, which are analogous to components of natural selection. These components are described below.

A. Encoding

Encoding is a way of representing the decision variables of the optimization problem in a string of binary digits called chromosomes. If there are v decision variables in an optimization problem and each decision variable is encoded as an n-digit binary number, then a chromosome is a string of $n\times v$ binary digits. Each chromosome is a possible solution to the optimization problem.

B. Initial population

The initial population is a set of chromosomes offered as an initial solution or as a starting point in the search for better chromosomes. The initial population must be large and diverse enough to allow evolution toward better individuals. In general, the population is initialized at random to a bit string of 0's and 1's. However, more directed methods for finding the initial population can sometimes be used to improve convergence time.

C. Fitness function

The fitness function rates chromosomes (i.e., possible solutions) in terms of how good they are in solving the optimization problem. It thus plays the role of the environment. The fitness function returns a single value for each chromosome, which is then used to determine the probability that this chromosome will be selected as a parent to generate new chromosomes. The fitness function is the primary GA component in which a traditional GA is tailored to a specific problem.

D. Genetic operators

Genetic operators are applied probabilistically to chromosomes of a generation to produce a new generation of chromosomes. Three basic operators are parent selection, crossover, and mutation. The parent selection operation mimics the natural selection process by selecting which chromosomes will be used to create a new generation, where the fittest chromosomes reproduce most often. The crossover operation refers to the exchange of substrings of two chromosomes to generate two new offspring. After parents are selected, and crossover generates two new chromosomes, the operation of mutation is applied to each bit in the string. Mutation simply alters the binary value of the bit when a random value generated for the bit is less than a predefined mutation rate.

E. Working parameters

A set of parameters, which includes the number of chromosomes in each generation, the crossover rate, the mutation rate, and the stopping criterion, is predefined to guide the GA. The crossover and mutation rates, assigned as real numbers between 0 and 1, are used as thresholds to determine whether the operators will be applied or not. The stopping criterion is predefined as the number of generations the algorithm is to be run or as a tolerance value for the fitness function.

Two forces, exploration and exploitation, interact in the search for better-fit chromosomes. Exploitation occurs in the form of parent selection. Chromosomes with higher fitness exploit this fitness by reproducing more often. Exploration occurs in the form of mutation and crossover, which allow the offspring to achieve a higher fitness than their parents. Crossover is the key to exploration, whereas mutation provides background variation and occasionally introduces beneficial genes into the chromosomes. For a successful GA, exploration and exploitation have to be in good balance. With too much exploitation, the GA may be stuck with copies of the same chromosome after a few generations, whereas with too much exploration, good genes may never be able to accumulate in the genetic pool.

GAs are ideal for sampling large search spaces and locating the regions of enhanced opportunity. Although GAs yield near-optimal solutions rather than optimal ones, obtaining such near-optimal solutions are usually the best that one can do in many complex optimization problems involving large numbers of parameters.

III. METHODS

A. Data set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsy in the Department of Radiology at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. To avoid the effect of repetitive grid lines on the image texture, mammograms that contained these grid lines caused by the stationary grid of some older mammographic units were excluded. The data set included 168 mammograms, with a mixture of benign ($n=85$) and malignant ($n=83$) masses. The visibility of the masses was ranked by an experienced breast radiologist on a scale of 1 to 10, where a ranking of 1 corresponded to the most visible category. The distribution of the visibility ranking of the masses is shown in Fig. 1. It can be observed that the visibility of the masses in our data set ranged from subtle to obvious.

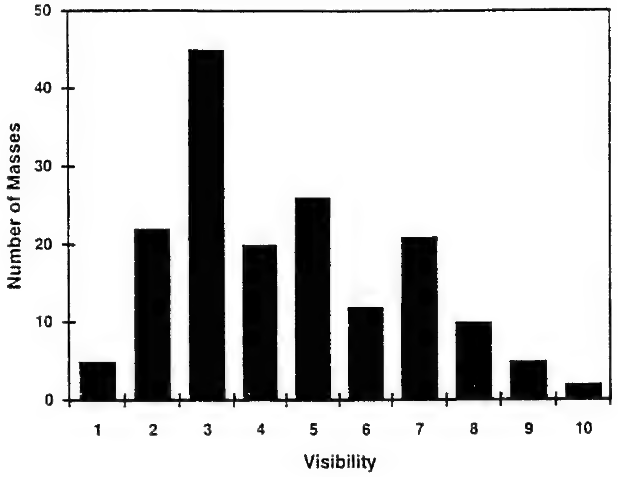


FIG. 1. The distribution of the visibility ranking of the masses in the data set.

The mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm×100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1- to 2.8-OD units, with a slope of -0.001-OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0 to 3.5.

Four different ROIs, each with 256×256 pixels, were selected from each mammogram. One of the selected ROIs contained the true mass as identified by an experienced radiologist and verified by biopsy. In addition to the ROI that contained the true mass location, the radiologist in the study was asked to select three presumably normal ROIs from the mammogram. The first of these three ROIs contained primarily dense tissue which could mimic a mass lesion, the second ROI contained mixed dense/fatty tissue, and the third contained mainly fatty tissue. An example of each of these ROIs is shown in Fig. 2.

B. Background correction

Breast masses are superimposed on structured background tissue in the ROIs. In most cases, this background tissue is not uniform over our 256×256 pixel ROI. For example, one side of the ROI may contain denser tissue than the other side, or, when the mass is close to the outer edge of the breast, one corner of the ROI may contain a nonbreast region. This nonuniformity may affect texture and morphological features that are extracted from the ROI. To reduce this effect, we developed a correction method that estimated the low-frequency background level based on the image intensities in a band of pixels surrounding the ROI. The background level at each pixel on the edge of the ROI was first estimated by gray-level averaging in a rectangular region surrounding the pixel. The background level of a pixel inside the ROI was then estimated by interpolation using the background pixel

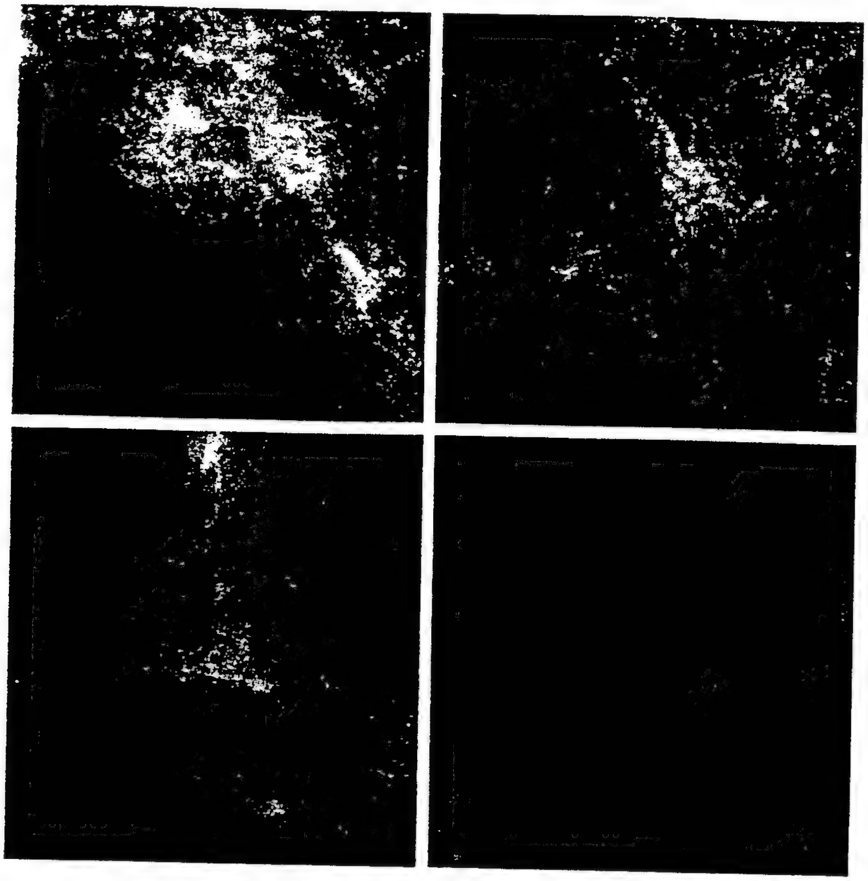


FIG. 2. An example of the mass and normal ROIs selected from one of the mammograms used in this study. The four ROIs are upper left—mass; upper right—mixed dense/fatty tissue; lower left—dense tissue; lower right—fatty tissue.

values on the edges. A more detailed description of this background correction method can be found in the literature.[10,32]

C. Feature extraction

1. Texture features

The texture features used in this study were calculated from spatial gray-level dependence (SGLD) matrices. The (i,j)th element of an SGLD matrix is the joint probability that gray levels i and j occur in a direction θ at a distance of d pixels apart in an image. We computed global texture features, which represent the average texture measures throughout the entire ROI, and local texture features, which represent (i) the texture measure of a denser subregion inside the ROI which is likely to contain the mass, and (ii) the texture difference between this subregion and other peripheral regions in the ROI which contain normal breast tissue. The method used for the computation of SGLD matrices and multiresolution texture analysis are explained in full detail elsewhere.[30] A brief description is given below.

Wavelet transform[33] using the four-coefficient Daubechies wavelet filter was applied to each ROI to decompose the image into a low-pass image and three high-pass subband images. For extracting global multiresolution texture features, we used the original image (scale$=1$) and the low-pass images at scales 2 and 4 to formulate SGLD matrices at $d=1$ in the transformed images. The distance of $d=1$ at these scales was equivalent to distances of 1, 2, and 4 in the original image. The wavelet coefficients at scale 8 were obtained with wavelet filtering but without down-sampling. The coefficients at scale 8 were used to formulate SGLD matrices at $d=2,3,4,\ldots,12$. Since no down-sampling was used at scale 8, these distances between pixel pairs were equivalent to distances of 8,12,16,...,48 in the original image. Thus a total of 14 distances were used. At each distance, four SGLD matrices at $\theta=0°$, 45°, 90°, and 135° were determined. Thirteen texture features were calculated from each SGLD matrix. The features at $\theta=0°$, 90° and at $\theta=45°$, 135° were averaged separately. Thus 26 texture features were computed for each d, resulting in a total of 364 global features.

For extracting local texture features, five subregions were automatically identified in the background-corrected ROI: a 90×90 pixel object subregion that contained the suspicious dense tissue or the mass, and four 64×64 pixel peripheral subregions that were located in the four corners of the ROI. The suspicious object subregion was automatically detected by searching for the highest average gray-level inside the ROI using a 90×90 moving box. For a given d, an SGLD matrix was derived from the object subregion, and a background SGLD matrix was derived from the pixel pairs in the

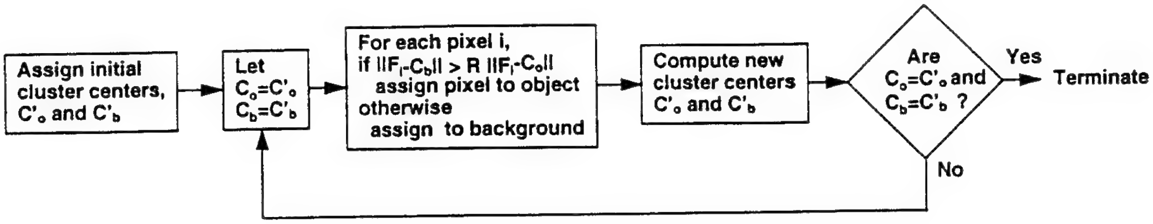

Assign initial cluster centers, C'_o and C'_b
Let $C_o=C'_o$ $C_b=C'_b$
For each pixel i, if $\|F_i-C_b\| > R \|F_i-C_o\|$ assign pixel to object otherwise assign to background
Compute new cluster centers C'_o and C'_b
Are $C_o=C'_o$ and $C_b=C'_b$?
Yes
Terminate
No

FIG. 3. A schematic of the clustering algorithm.

four peripheral subregions. The SGLD matrices were computed at $d=1,2,4$, and 8. Analogous to the global texture feature extraction, for a given d, 26 features were computed from the object SGLD matrix, and 26 features were computed from the background SGLD matrix. The local texture feature space therefore consisted of 104 features extracted from the object subregion, and 104 features calculated as the differences between the corresponding features extracted from the object and peripheral subregions. This resulted in a total of 208 local texture features.

The detail images in the wavelet transform can be expected to contain useful information for texture-based classification of a large class of images. However, in our previous studies, we found that using the SGLD texture features based on the detail images in the wavelet transform domain did not result in proper classification of breast masses and normal breast tissue.[11] Since this study focused on the feature selection aspect of classification, we did not attempt to search for new texture features that are presumably present in the detail images.

2. Morphological features

We have developed an automated algorithm for segmentation of an ROI into an object region and background tissue.[31] The morphological features are extracted automatically from the object region after the segmentation is performed.

We used a pixel-by-pixel clustering algorithm followed by binary object detection for ROI segmentation. Pixel-by-pixel clustering algorithms have found widespread use in segmentation of remote sensing data,[34] where multispectral and/or multisource data are obtained for each pixel in the image. Data points for each pixel are regarded as components of a multidimensional feature vector, and pixels with feature vectors of similar characteristics are assigned to the same class using a clustering algorithm. Our data set contains a single data point (the gray level) for each pixel. We derived several filtered images from this single image, and used the original and filtered pixel values as the components of the feature vectors in the clustering algorithm. Inclusion of the filtered images makes it possible to incorporate neighborhood information into the classification of each pixel.

Our clustering algorithm, depicted in Fig. 3, is very similar to the migrating means algorithm.[34] The goal is to classify pixel p_i as either an object or a background pixel. This is achieved by clustering with feature vector $F_i=[f(1),...,f(L)]$ of length L, where L is the total number of images used in clustering. The algorithm starts by choosing initial cluster center vectors, for the object and the background, as described below. Let $C_o=[c_o(1),...,c_o(L)]$ and $C_b=[c_b(1),...,c_b(L)]$ denote these cluster center vectors, respectively. Let $d_o(i)$ denote the Euclidean distance between F_i and C_o .$d_b(i)$ denote the Euclidean distance between F_i and C_b, and R denote a constant distance ratio. If $d_b(i)/d_o(i)>R$, the pixel p_i is temporarily classified as an object pixel; otherwise, it is classified as a background pixel. If $R=1$, the algorithm becomes identical to the migrating means algorithm. After this temporary classification, two new cluster center vectors are computed. The lth component of the new object and background center vectors are the averages of the lth components for pixels temporarily classified as object and background pixels, respectively. If the new cluster centers are different from the previous ones, the procedure of temporary classification is repeated, otherwise, the clustering is completed. In this paper, we used $R=3.75$ so that F_i had to be much closer to C_o than to C_b to be classified as an object pixel. This conservative criterion reduces the chance that a mass region merges with adjacent tissue. However, it also slightly underestimates the mass size so that the detected edge is often within the margin of the mass. The initial center vectors were chosen such that each component of the initial object center vector is 1.1 times the average of that component over the entire ROI, and each component of the initial background center vector is 0.9 times the same average.

After clustering, the ROI may contain several disconnected objects. To obtain a single suspected mass object, we selected the largest connected object among all detected objects. We finally applied region growing to a small region outside the boundary of the suspected object to get a better definition of its borders. To achieve this, we thresholded the original image pixels that were within ten pixels of the object border. The threshold value was chosen experimentally to be the difference between the mean of the pixel values inside the object and half of their standard deviation. Figure 4 shows an example of the result of our segmentation algorithm.

In this paper, we used three filtered images along with the original image to form the feature vectors. The first filtered image was obtained by median filtering with a 5×5 kernel. The second and third filtered images were edge-enhanced images at different resolutions.[31] Each filtered image, as well as the original image was linearly normalized between 0 and S_l, where S_l, $l=1...L$ is a scaling factor. The scaling factors

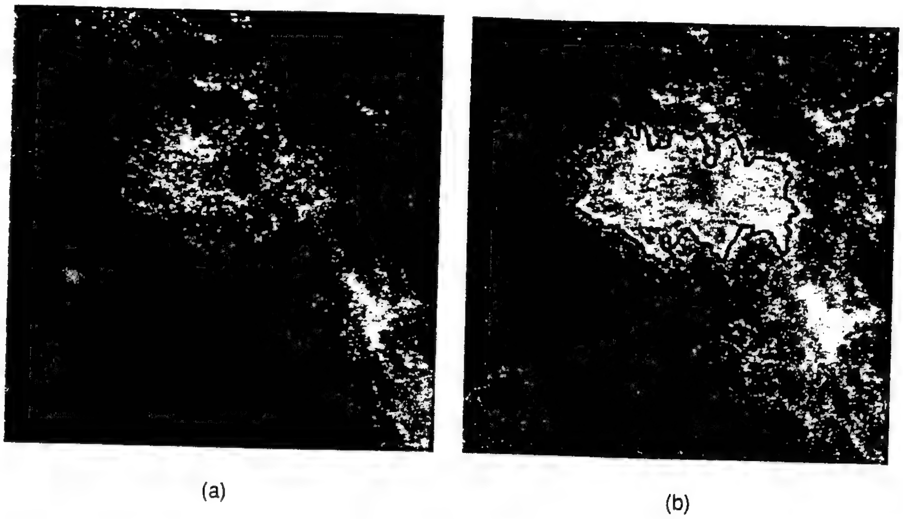


FIG. 4. (a) An ROI with an ill-defined mass, (b) mass object extracted automatically by the clustering algorithm and superimposed on the background-corrected ROI.

S_l were chosen experimentally to be $S_1=S_2=1400$ for the original image and the median filtered image, and $S_3=S_4=770$ for the edge-enhanced images. Therefore, the original image and the median filtered image were weighted approximately twice as much as the edge-enhanced images in the clustering algorithm. This bias in favor of the weights of the original image and the median filtered image was necessary because the algorithm showed a tendency to segment only disconnected edges if all images were equally weighted.

After detection of a single suspicious object within each ROI, features were extracted from the object and its margins. We extracted eleven shape features from each object, and four features from the margins of each object. The shape features included the number of edge pixels, area, circularity, rectangularity, contrast, the ratio of the number of edge pixels to the area, and five normalized radial length features. A detailed discussion of the shape features used in this study can be found in Ref. 35. The margin features were computed as follows. First, the mean and the standard deviation of the pixel values inside the object were computed. Next, pixels in a boundary region outside the object but within a distance of 15 pixels from the object border were thresholded. The values of the thresholds were chosen to be the mean minus 0.5, 1, 1.5, and 2 times the standard deviation. The number of pixels in the boundary region which were above the thresholds was defined as the margin features. Thus a total of 15 morphological features were extracted from each ROI.

D. Classifiers

In this paper, we investigated GA-based feature selection for two kinds of classifiers, namely (i) a linear classifier based on Fisher's linear discriminant;[20] and (ii) a multilayer backpropagation neural network (BPN).[36] For each ROI, both classifiers produced a scalar, termed the classifier output, which indicated the likelihood that the ROI contained a real mass.

Fisher's linear discriminant is based on a linear projection of the feature space onto the real line such that the ratio of the between-class sum of squares to within-class sum of squares is maximized after the projection.[20] In our two-class problem, the statistical procedure for formulation of the linear discriminant function is equivalent to multiple linear regression.[37] Fisher's linear discriminant is the optimal classifier if the features are distributed as multivariate Gaussian random variables with equal covariance matrices under each class.[21]

The BPN used in this study consisted of an input layer, an output layer, and a single hidden layer. Each layer in the BPN contained a number of nodes, which were connected to previous and subsequent layers by trainable weights. A single feature was applied to each node in the input layer. The net input to each node in the hidden layer and the output layer was a weighted sum of the node outputs from the previous layer. The output of a node was related to its net input by a sigmoidal function. The output layer contained a single node, whose output indicated the likelihood that the ROI contained breast mass tissue. The BPN was trained using batch processing and the delta-bar-delta rule for improved rate of convergence and stability.[32]

Since our purpose in this study is to design a feature selection algorithm, we did not compare BPN and linear discriminant classifiers. Instead, we compared the classification accuracy obtained by using different feature selection methods, with a fixed classifier for each comparison.

E. GA-based feature selection

In this paper, we used a GA to select features for discrimination of mass and nonmass ROIs. In our GA, the number of bits in a chromosome was equal to the total number of available features, and each bit corresponded to an individual feature extracted from the ROIs. A feature was termed "present" in a chromosome if the value of the bit corresponding to that feature was 1. The population was initialized at random, with a small probability P_{init} of having a 1 at each bit location. This allowed the GA to start with a few selected features and grow to a reasonable number of features as the population evolved. The total number of chromosomes at each generation was kept constant at $M=250$.

At each run of the GA, the image data set of 672 ROIs was divided into a training and a test set, with ROIs belonging to the same film grouped into the same set. The training set was used in the GA for feature selection. After feature selection, a classifier was trained using only the GA-selected features of the training set. The classification accuracy of the procedure was evaluated by applying the classifier to the same set of features of the test group, as described below. For studying the effect of GA parameters on the classification accuracy with the linear discriminant classifier, ten random partitionings of training and test sets were obtained for each set of different GA parameters, and the results were averaged in order to reduce the effect of case selection. For experiments with the BPN, 50 random partitionings were used. For both experiments, the number of mass and nonmass ROIs in each training set was 126 and 378 ($\frac{3}{4}$ of the total), respectively, while the number of mass and nonmass ROIs in each test set was 42 and 126 ($\frac{1}{4}$ of the total), respectively.

Inside the GA, the training set was equally divided into two groups, $S1$ and $S2$. For each chromosome, two classifiers were trained, with $S1$ and $S2$ as the training groups, respectively. Only the features present in the chromosome were used as features in classifier training. The classifier trained on group $S1$ was applied to the group $S2$, and vice versa, for calculation of two sets of *pseudotest* classifier outputs. The accuracy of the pseudotest classifier outputs, and the number of selected features were then used to define the fitness of the individual chromosome. This process was repeated for each of the M chromosomes in each generation.

The main component of the fitness function was the area A_z under the receiver operating characteristic (ROC) curve of the pseudotest sets. A widely accepted procedure for computing the ROC curve assumes that the classifier output follows a normal distribution for each class, and fits the ROC curve to the classifier output using maximum likelihood estimation.[38] We adopted this approach when we studied and compared the classification accuracy of our classifiers with the selected feature sets. However, it is computationally expensive to use this approach in the fitness function calculation inside the GA, because it is required for each chromosome in each generation. Instead, we chose to estimate the ROC curve by varying the decision threshold, and determining the true-positive fraction (TPF) as a function of the false-positive fraction (FPF). The A_z value was estimated by numerical integration using the trapezoidal rule. Since the estimation of the A_z was internal to the GA, it did not affect the A_z values reported in the Sec. IV for a set of selected features. Internal to the GA, the fitness ranking of the chromosomes might be slightly different from that obtained by using the maximum likelihood ROC curve. However, the effect on the final selected feature set should be small, because this slight difference did not completely eliminate the lower-ranking chromosomes. A slightly lower-ranking chromosome was assigned a slightly lower probability of being a parent, but it could still be competitive after mutation and crossover if it contained effective features. This minor inaccuracy in the fitness function computation was a trade-off in order to execute the computation in a reasonable amount of time while using the A_z value in the feature selection procedure.

A second component of the fitness function was a penalty term, analogous to Brill's utility term,[28] which was linearly proportional to the number of features present in the chromosome. The purpose of this penalty term was to control the number of selected features and to prevent overfitting in the test stage of classifier design. In other words, the penalty term was designed to improve the classification accuracy, and not for accelerating the computational speed. The function of the penalty term was comparable to those of the F-to-enter and F-to-remove thresholds in the stepwise feature selection method, described in the next subsection. Similar to these corresponding parameters in stepwise feature selection, increasing the penalty term decreased the number of selected features. We studied the effect of the presence of this penalty term on the test results.

In a given generation, the fitness function $f(m)$ for a chromosome m was computed as follows. First, the two pseudotest A_z values, corresponding to pseudotest sets $S1$ and $S2$, were averaged to yield $\bar{A}_z(m)$. Next, a fitness function $\bar{f}(m)$ was computed as

$$\bar{f}(m)=\bar{A}_z(m)-\alpha N(m), \tag{1}$$

where $N(m)$ was the number of 1's (present features) in chromosome m and α was the penalty constant. After $\bar{f}(m)$ was determined for all chromosomes, the maximum $\bar{f}_{\max}$ and the minimum $\bar{f}_{\min}$ of $\bar{f}(m)$ over the population of M chromosomes were calculated. Finally, $\bar{f}(m)$ was normalized using $\bar{f}_{\max}$ and $\bar{f}_{\min}$ to yield the fitness function $f(m)$,

$$f(m)=\left(\frac{\bar{f}(m)-\bar{f}_{\min}}{\bar{f}_{\max}-\bar{f}_{\min}}\right)^2, \quad 1\leq m\leq M. \tag{2}$$

The genetic operators were applied as follows. First, parent selection was performed using roulette wheel selection.[23] In this method, each chromosome in a generation occupies an area

$$A(m)=\frac{f(m)}{\sum_{m=1}^{M} f(m)} \tag{3}$$

proportional to its fitness, on a roulette wheel. A parent is selected by spinning the roulette wheel, i.e., by generating a random number $\gamma_i \in (0,1]$ and determining the chromosome m_i that satisfies

$$\sum_{m=1}^{m_i-1} A(m)<\gamma_i\leq \sum_{m=1}^{m_i} A(m), \quad i=1,2. \tag{4}$$

After two parents m_1 and m_2 were selected for generating two offspring, a probabilistic decision was made as to whether crossover should be applied or not. A random number β with uniform distribution in the interval (0,1] was generated and compared to P_c, the probability of crossover. If $\beta > P_c$, then no crossover was applied, and m_1 and m_2 were accepted into the new generation. Otherwise, a random crossover site was selected inside the chromosomes, and each of the parent chromosomes were split into left and right strings at this location. Crossover was completed by combining the left string of m_1 with the right string of m_2, and vice versa.

Finally, mutation was applied to each bit of the chromosomes in the new generation. Again, a random number with uniform distribution in the interval (0,1] was generated, and compared to P_m, the probability of mutation. If P_m was higher, then the bit was complemented. Otherwise, it was left unchanged. We studied the effects of P_c and P_m on the final classification accuracy.

The GA was permitted to evolve for a fixed number of generations. After the evolution was completed, the chromosome with the highest fitness value provided the set of selected features. The entire training set $S1 \cup S2$ was then used in the final multiple linear regression to determine the weight of each selected feature in the classifier. During testing, the values of the selected features of each ROI in the test set were applied as inputs to the trained classifier to calculate the classifier output for that ROI.

To evaluate the classification performance, the classifier output was used as the decision variable, and a test ROC curve was estimated using the LABROC1 program.[39] The LABROC1 program assumes binormal distributions of the decision variable for the normal and abnormal cases, and fits the ROC curve based on maximum likelihood estimation. The area under the fitted ROC curve, A_z, was used as an index of classification accuracy.

F. Stepwise feature selection

For the purpose of comparison with GA-based feature selection, we also studied the classification accuracy of the same classifiers using a well-established feature selection method, called feature selection with stepwise linear discriminant analysis,[21] or stepwise feature selection in short.[40] At each step of the stepwise selection procedure, one feature is entered into or removed from the selected feature pool by analyzing its effect on a selection criterion. In this study, we employed the Wilks' lambda as our selection criterion, which is defined as the ratio of the within-group sum of squares to the total sum of squares of the two classes.[37] The number of features selected by this method are controlled by two parameters, called F-to-enter and F-to-remove. At each step, the stepwise feature selection algorithm first determines the significance of the change, based on F statistics, in Wilks' lambda when a variable is entered into the selected feature pool. If the significance is above the threshold determined by the F-to-enter parameter, then the selected feature pool is augmented with the most significant variable. Next, the algorithm computes the significance of the change in Wilks' lambda when each variable is removed from the selected feature pool. If the significance is below the threshold determined by the F-to-remove parameter, then the least significant variable is removed from the selected feature pool. Increasing either the F-to-enter or the F-to-remove value decreases the number of selected features. Similar to GA-based feature selection, stepwise feature selection is a heuristic procedure. For this reason, the optimal values of F-to-enter and F-to-remove parameters are not known in advance. One has to experiment with these parameters and increase or decrease the number of selected features to obtain the best test performance. A detailed description of the stepwise feature selection procedure and its application to our problems[10,11] can be found in the literature.[21,40]

IV. RESULTS

In the next two subsections, we present the results for evaluation of the effects of various parameters, and for classification with GA-based feature selection using linear discriminant and BPN classifiers, respectively. Since training a linear discriminant classifier was considerably faster than training a BPN, the effects of GA parameters were studied with a linear discriminant classifier. Feature selection for a BPN classifier was performed on a subset of the entire feature set to accelerate training. For both classifiers, a comparison with stepwise feature selection was provided.

A. Feature selection for a linear discriminant classifier

1. Effect of penalty term and number of generations

To determine a reasonable number of generations for the GA to evolve, we selected several combinations of crossover probability (P_c) and mutation probability (P_m), and monitored the growth of the number of selected features. The initial probability of feature presence was fixed at $P_{\text{init}}=0.002$. The GA was allowed to evolve with two different α values of the penalty term in the fitness function of Eq. (1). We observed that the crossover probability P_c did not have a major effect on the number of selected features. However, both α in the penalty term and the mutation probability P_m affected the number of selected features. Figures 5 and 6 plot the average number of selected features over ten training sets versus the generation number for $\alpha=0$ and $\alpha=1/2000$, respectively. The average number of selected features is plotted for $P_m=0.001$ and $P_m=0.003$ in each figure. The crossover probability is kept constant at $P_c=0.7$. The test A_z value obtained up to a given generation is plotted against the generation number in Figs. 7 and 8 for the same conditions ($\alpha=0$ and $\alpha=1/2000$), respectively. The average A_z value over ten test sets is shown.

It is observed that while the average test A_z value does not increase after the 25th generation, the number of selected features keeps increasing beyond the 60th generation for all combinations of GA parameters studied. Since the main component of the fitness function in the GA is the A_z value rather than the number of features, more features may be added into the selected feature pool as long as the area under the ROC curve does not deteriorate. Comparing Figs. 5 and

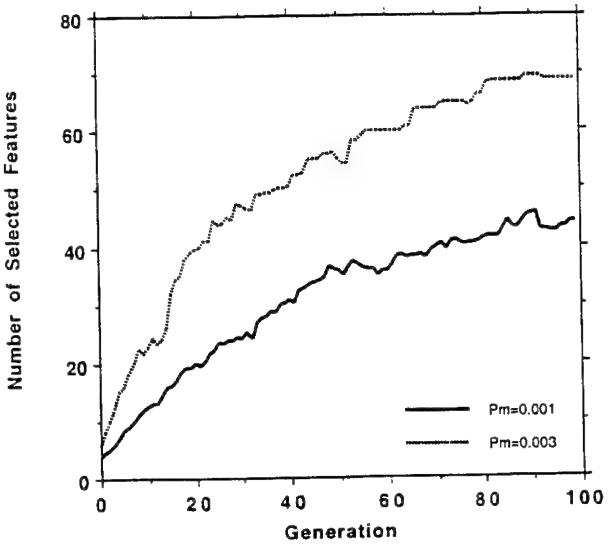


FIG. 5. Evolution of the number of selected features for α=0.

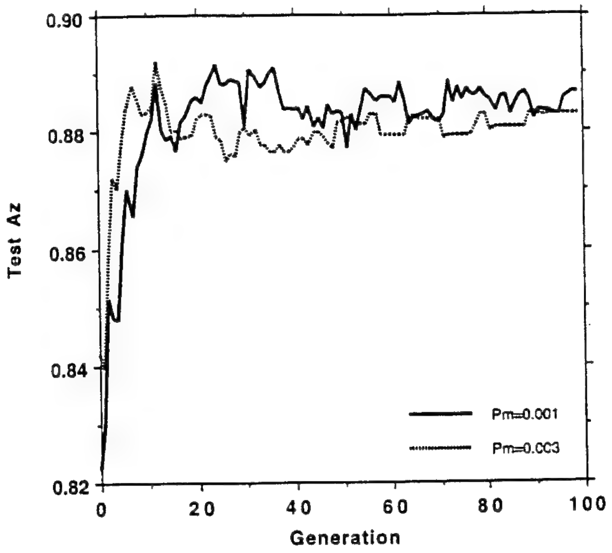


FIG. 7. Evolution of the average test A_z for α=0.

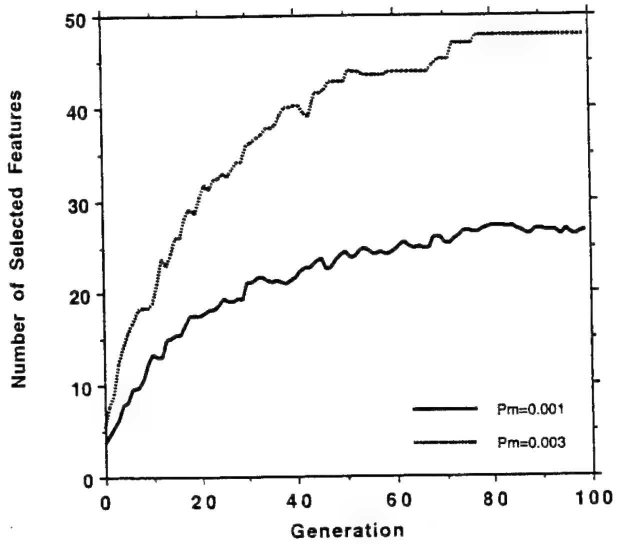


FIG. 6. Evolution of the number of selected features for α=1/2000.

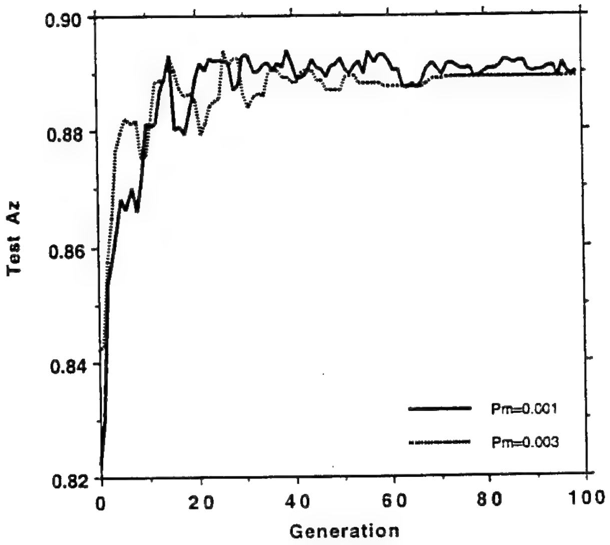


FIG. 8. Evolution of the average test A_z for α=1/2000.

6, it can be observed that the penalty term suppressed the number of selected features. The number of selected features eventually leveled off at about the 80th generation when the penalty term was nonzero (Fig. 6).

The average test A_z values at the end of 100 generations were 0.89 for the combinations studied in Fig. 8, and 0.88 for the combinations studied in Fig. 7. The maximum and minimum values of individual test scores for the ten partitions studied were 0.92 and 0.86 for Fig. 8, and 0.92 and 0.85 for Fig. 7. The standard deviation of the individual A_z values, as determined by the LABROC1 program, varied between 0.02 and 0.04.

Since our goal is to select a small number of features while maintaining a high classification accuracy, we performed subsequent GA experiments with α=1/2000. Due to computation time constraints, we set the maximum number of generations to be 25 in the following experiments.

2. Effect of initial probability of feature presence (P_{init})

We evaluated the effect of P_{init} on feature selection when the crossover probability P_c and the mutation probability P_m were held constant. The average test A_z values for P_c=0.9 and P_m=0.001 are tabulated in Table I. It is observed that the performance of the GA reaches a broad maximum when P_{init} is in the range of 0.0005 to 0.020, i.e., when the average number of features in the initial chromosomes is approximately in the range of 0.3 to 12. When P_{init} is out of this range, the average test A_z decreases slightly.

3. Effect of probability of mutation and crossover

The effects of the crossover probability P_c and the mutation probability P_m on the classification accuracy are summarized in Tables II and III, respectively. In Table II, the

TABLE I. The effect of P_{init} on GA performance for P_m=0.001, P_c=0.9.

P_{init}	Average test A_z	Avg. Num. of features
0	0.88	18.5
0.0005	0.89	17.2
0.001	0.88	20.8
0.002	0.90	20.1
0.005	0.89	18.0
0.010	0.89	23.2
0.020	0.89	22.9
0.050	0.88	32.7

control parameters were fixed at P_{init}=0.002, and P_m=0.001, while in Table III, they were fixed at P_{init}=0.002, and P_c=0.9. For fixed values of P_{init} and P_m, the average test A_z appears to increase with increasing P_c, while the number of selected features remains relatively constant. On the other hand, for fixed values of P_{init} and P_c, the average test A_z increases initially with increasing P_m, reaching a maximum at P_m=0.001, and then decreases slightly as P_m increases beyond 0.003. Although the variation of the classification accuracy with respect to P_m is not significant, it appears that a reasonable range of choice for P_m is such that the average number of mutations per chromosome per generation is less than 1.5 (0.003× the number of genes per chromosome). Within the range studied, the number of selected features increases with increasing P_m, which may be the reason for the slight deterioration in performance for large P_m.

4. Comparison with LDA classifier and random feature selection

We used a commercial statistics package, SPSS,[40] for LDA classification. The feature selection and formulation of the discriminant function were performed on each of the ten training sets, and the discriminant functions were tested on the corresponding test sets. Using minimization of Wilks' lambda as the feature selection criterion, we varied the two threshold values for F statistics (F-to-enter and F-to-remove) in the SPSS package so that the average test A_z value over the ten partitionings was maximized. The number of selected features and the test results for the ten partitionings are tabulated in Table IV. We chose the best GA classification results (the last line in Table II) for comparison with those of the LDA. The corresponding test A_z values and the number of selected features for each partitioning of the data set are tabulated in Table IV.

TABLE II. The effect of P_c on GA performance for P_{init}=0.002, P_m=0.001.

P_c	Average test A_z	Avg. Num. of features
0.1	0.87	18.4
0.3	0.89	18.6
0.5	0.89	17.8
0.7	0.89	18.3
0.9	0.90	20.1

TABLE III. The effect of P_m on GA performance for P_{init}=0.002, P_c=0.9.

P_m	Average Test A_z	Avg. Num. of features
0.0005	0.89	16.0
0.001	0.90	20.1
0.003	0.89	32.3
0.005	0.88	33.1
0.007	0.89	33.4
0.009	0.88	33.9

For comparison with these two near-optimal feature selection methods, we performed multiple linear regression training and testing on 20 randomly selected features out of the available 587 features. The test A_z values based on these 20 randomly selected features are also given in Table IV.

B. Feature selection for BPN

Since training a BPN is considerably slower than training a linear discriminant classifier, we modified our training strategy for this classifier. The basic differences between the experiments in this subsection on BPN and the previous subsection on linear discriminant classifier were: (1) In order to handle a smaller feature pool with BPN, we used a single distance for texture features. Based on our previous study of the effects of pixel distance on classification,[10] we selected a pixel distance of d=20. The global texture features computed at this pixel distance, plus the morphological features previously described in Sec. III C, constituted the feature pool in this subsection. Therefore, there were a total of 41 features (26 texture and 15 morphological) for the feature selection algorithms to choose from. (2) In order not to repeat the feature selection process several times with several different training sets, the entire data set was used in the feature selection step of the classification procedure. After feature selection was completed, the classifier was trained and tested with 50 different partitionings of the data set into training and test groups. As in the case of linear discriminant classifier, the number of mass and nonmass ROIs in each training set was 126 and 378 ($\frac{3}{4}$ of the total), respectively, while the number of mass and nonmass ROIs in each test set was 42 and 126 ($\frac{1}{4}$ of the total), respectively.

The parameters of the BPN and the GA used in this subsection were as follows. The BPN had a variable number of input nodes, four hidden layer nodes, and a single output node. The BPN was trained for 400 iterations for each chromosome in each generation. The GA was allowed to evolve for a total number of 75 generations. Results of the previous subsections suggest that there is a wide range of choice for the parameters P_{init} and P_m. It appears that a reasonable choice for P_{init} is such that the average number of selected features at generation 0 is in the range of 0.3 to 12, and a reasonable choice for P_m is such that the average number of mutations per chromosome per generation is less than 1.5. For this reason, these parameters of the GA were selected as P_m=0.02, and P_{init}=0.02. Since a large probability of crossover seemed to result in the selection of more effective fea-

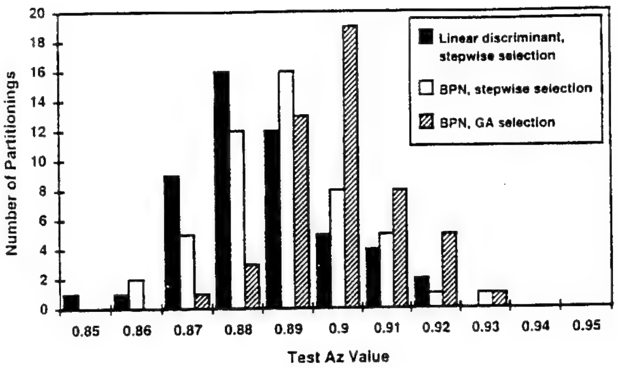


FIG. 9. The distribution of the test A_z values for the linear classifier with stepwise feature selection, BPN classifier with stepwise feature selection, and BPN classifier with GA feature selection.

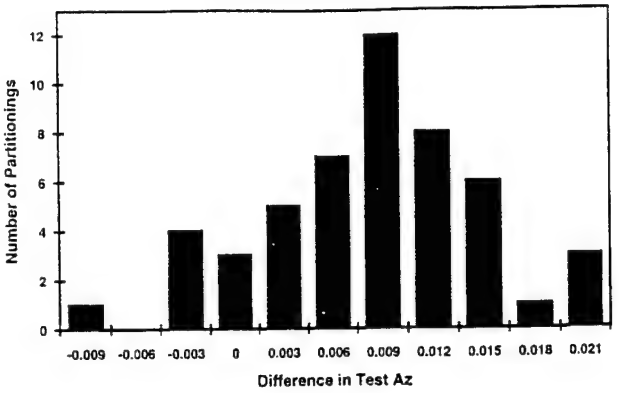


FIG. 10. The distribution of the pairwise difference of the test A_z values of the BPN classifiers with GA and stepwise feature selection.

tures in the previous subsections, the value of P_c was chosen as 0.9. A penalty term was applied to the fitness function with $\alpha=1/2000$.

The final GA-selected pool of variables contained 16 features. After feature selection using the GA, the performance of the BPN classifier with the selected features was tested with 50 training and test groups as described above. The average training and test A_z values over 50 partitionings were 0.92 and 0.90, respectively.

To compare our GA-based feature selection method for a BPN, we also used the same data set and the 41 features described above with stepwise feature selection. The entire data set was used for feature selection. The final selected pool of variables contained 19 features. The same 50 partitionings used for the GA experiments were used to train and test both a linear discriminant classifier and a BPN with the stepwise-selected features. The average training and test A_z values over 50 partitionings were 0.92 and 0.89 with the linear classifier, and 0.92 and 0.89 with the BPN classifier. The distribution of the test A_z values for the linear classifier, as well as the BPN classifier with features selected using stepwise and the GA-based feature selection are shown in Fig. 9. The distribution of the pairwise difference of the test A_z of the BPN classifiers with stepwise and GA-based feature selection methods is shown in Fig. 10.

V. DISCUSSION

Our goal in this paper was the development of an effective feature selection algorithm given a large number of features extracted from an image data set. Table IV and Figs. 9 and 10 indicate that GA feature selection might be a viable alternative to stepwise feature selection.

The average number of features selected by stepwise and GA-based feature selection methods for a linear discriminant classifier were 19.3 and 20.1, respectively, in Table IV. In the same table, we compared these methods to random feature selection with the number of selected features equal to 20. Both methods performed better than random feature selection. The difference between the average A_z obtained by GA-based feature selection and random feature selection was more than two times the standard deviation of each A_z distribution.

We observed that each time the GA was trained with a different training set, a different set of features was selected. This was also true for stepwise feature selection. The basic reason for this was the limited size of the data set. If training sets that could represent the entire population were available, the selected set of features could be expected to be more consistent among different training sets. With the limited data set used in this study, each time a set of cases was left out as the test data, the statistical characteristics of the training feature set changed. Furthermore, many of the features were highly correlated, with correlation coefficients close to 1 or −1. Therefore, these correlated features could be interchanged. Only ten features were selected three or more times for the experiments in Table IV. Out of these ten features, six were texture and four were morphological features. This indicates that morphological and texture features are both important for the classification of the ROIs.

The high correlation between the features in the feature space used in this study is probably a cause of the surprisingly good classification result (A_z=0.82) obtained with the randomly selected features. This may also indicate that many of the features in the feature space are very effective for this

TABLE IV. Test A_z values of a linear discriminant classifier using stepwise LDA, GA-based feature selection, and 20 randomly selected features.

	Stepwise LDA		GA		Random
Test group	A_z	Num. of features	A_z	Num. of features	A_z
1	0.87	19	0.90	20	0.80
2	0.91	15	0.89	24	0.86
3	0.92	25	0.93	24	0.86
4	0.88	22	0.88	20	0.81
5	0.86	23	0.84	23	0.78
6	0.92	19	0.93	20	0.83
7	0.92	15	0.91	17	0.87
8	0.84	21	0.88	19	0.75
9	0.86	14	0.88	18	0.77
10	0.88	20	0.92	16	0.82
Average	0.89	19.3	0.90	20.1	0.82

classification task. Therefore, even when only 20 features are randomly drawn, we have a high probability of drawing effective features and obtaining a classification result that is much higher than that would be obtained by chance.

Our results indicate that the classification results with GA-based feature selection are better than their counterparts with stepwise feature selection. This is most easily seen from Fig. 9, which compares the distribution of the A_z values for a BPN classifier with GA-based feature selection to that with stepwise feature selection. It can be observed that the two distributions are shifted with respect to each other, with the distribution using GA-based feature selection exhibiting higher A_z values. However, we could not perform a paired t-test to evaluate the statistical significance of the differences for the results listed in Table IV or those shown in Fig. 9. The paired t-test requires independence among the samples whereas our test (or training) sets in the different partitionings overlapped with each other. We have used the CLABROC program[41] to test the statistical significance of the difference between the corresponding pair of ROC curves for each partitioning. The difference did not achieve statistical significance for the individual pairs because the number of cases in each partitioned data set is small and thus the standard deviation of A_z is large (0.02 to 0.04). However, it should be noted that the improvement in A_z with GA-based feature selection, although small, is consistently observed over the different partitionings of the data set, over both the linear discriminant classifier (Table IV) and the BPN classifier (Figs. 9 and 10), as well as over different data sets.[42] The small improvement in A_z may be attributed to two causes: (1) For the linear discriminant classifier, the stepwise feature selection procedure is already near optimal. It is actually somewhat unexpected that the GA-based feature selection can still provide an observable improvement in A_z. (2) It is well known that BPN performance may not reach the global maximum if there are insufficient training samples. For the BPN classifier in this study, the number of weights to be trained was large compared with the number of input training samples. Therefore, it probably did not reach its optimum when it was used in a GA for feature selection. Again, a consistent improvement in A_z demonstrates that the GA can select more effective features for BPN classifiers.

The main advantage of GA-based feature selection is its flexibility. GA-based feature selection can be applied to any classifier and the fitness function can be tailored to select features with specific characteristics. An example of the former application is to select features for a nonlinear classifier such as a BPN as discussed above. An example of the latter application is to select features for development of a highly sensitive classifier[43] described next.

In both breast cancer detection and classification, the cost of missing a malignant lesion is very high. For this reason, an important measure of classification accuracy is the FPF at high true-positive classification. Since the design of the fitness function of a GA is very flexible, one can target to maximize the partial area above a specified TPF in order to optimize the classifier performance in this region. In a preliminary study with our data set,[43] we designed a GA-based feature selection algorithm in which the fitness of a chromosome was defined as the partial area above a TPF of 0.95. We then compared the FPF at TPFs of 100% and 96% using GA-based and stepwise feature selection for a linear discriminant classifier. At a TFP of 100%, the average FPF over the ten partitionings used in this study were 0.44 for GA-based feature selection, and 0.68 for stepwise feature selection. At a TFP of 96%, the average FPFs were 0.33 for GA-based feature selection, and 0.38 for stepwise feature selection. These encouraging results demonstrate the potential of a GA-based approach to designing classifiers for a wide range of practical problems, which cannot be achieved with a conventional method such as stepwise discriminant analysis.

Stepwise feature selection is computationally faster than GA-based feature selection. For example, in the present study, the stepwise feature selection required 64-s CPU time for each partition (Table IV) on a 90-MHz Pentium-based personal computer. The GA-based feature selection required 519-s CPU time for each partition (Table IV) on a 133-MHz alpha-based workstation, when the evolution involved a total of 250 chromosomes. However, a GA is highly parallelizable. In principle, the fitness of each chromosome can be evaluated on a different processor and the computation time can be improved up to a factor equal to the number of chromosomes. The choice between GA-based or stepwise feature selection will depend on the application. For a linear discriminant classifier, the stepwise feature selection may be near optimal so that the advantage of using a GA may be small. However, for other classifiers, a GA may be more effective because the selected feature set will be optimized to the specific classifier used.

A GA was previously used for the task of feature selection in a classification problem with 30 features and 150 cases.[28] The GA fitness criterion in this application was designed to be a function of the correct classification rate with a nearest-neighbor classifier. After the features were selected, a neural network was employed for final classification. Our approach has two advantages over this application. First, we used a more sophisticated classifier in the fitness function computation stage, hence GA training is more efficient. Second, we used the same classifier at the final classification stage, therefore our results are expected to be more consistent. Our results are also expected to be less biased since we divided our data set into independent training and test groups for GA evaluation, whereas the entire data set was used for training in the other study.[28]

VI. CONCLUSION

We investigated the use of a GA for feature selection, and demonstrated its application by classifying ROIs on mammograms as either containing mass or normal tissue. By comparing stepwise feature selection and GA-based feature selection for two different classifiers (the linear discriminant classifier and the BPN), and by examining the problem of designing classifiers biased to have high sensitivity performance, we have demonstrated the versatility offered by a GA in the design of classifiers for a variety of classification tasks without a trade-off in the effectiveness of the selected features. Future work in this area includes application of GA-based feature selection to different classification tasks such as differentiation of malignant and benign tissue, and a detailed investigation of the formulation of different fitness measures, such as the partial area at the high-TPF region of the ROC curve, for the design of classifiers in different applications.

ACKNOWLEDGMENTS

This work is supported by the USPHS Grant No. CA 48129 and U.S. Army Grant No. DAMD 17-93-J-3007 (through subgrant No. GU RX 4300-803UM from Georgetown University). The content of this publication does not necessarily reflect the position of the Georgetown University or the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for providing the LABROC1 program.

[1] C. J. Vybomy, "Can computers help radiologists read mammograms?," Radiology **191**, 315–317 (1994).

[2] R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," Radiology **184**, 613–617 (1992).

[3] M. G. Wallis, M. T. Walsh, and J. R. Lee, "A review of false negative mammography in a symptomatic population," Clin. Radiol. **44**, 13–15 (1991).

[4] D. B. Kopans, "The positive predictive value of mammography," Am. J. Radiol. **158**, 521–526 (1991).

[5] D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," Curr. Opinion Radiol. **4**, 123–129 (1992).

[6] D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammography," IEEE Trans. Med. Imag. **9**, 233–241 (1990).

[7] F. F. Yin, M. L. Giger, C. J. Vybomy, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," Invest. Radiol. **28**, 473–481 (1993).

[8] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," IEEE Trans. Med. Imag. **12**, 664–669 (1993).

[9] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).

[10] H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).

[11] D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," Med. Phys. **22**, 1501–1513 (1995).

[12] I. E. Magnin, A. Bremond, F. Cluzeau, and C. L. Odet, "Mammographic texture analysis—An evaluation of risk for developing breast cancer," Opt. Eng. **25**, 780–784 (1986).

[13] Y. Wu, M. L. Giger, K. Doi, C. J. Vybomy, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," Radiology **187**, 81–87 (1993).

[14] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," Med. Phys. **19**, 1475–1481 (1992).

[15] Y. Chitre, A. P. Dhawan, and M. Moskowitz, "Artificial neural network based classification of mammographic microcalcifications using image structure features," Int. J. Pattern Recognition Artificial Intelligence **7**, 1377–1401 (1993).

[16] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem in digital chest radiograph segmentation," IEEE Trans. Med. Imag. **14**, 537–547 (1995).

[17] T. M. Cover and J. M. V. Campenhout, "On the possible orderings in the measurement selection problem," IEEE Trans. Syst. Man Cybern. **7**, 657–661 (1977).

[18] T. M. Cover, "The best two independent measurements are not the two best," IEEE Trans. Syst. Man Cybern. **4**, 116–117 (1974).

[19] W. S. Meisel, *Computer-Oriented Approaches to Pattern Recognition* (Academic, New York, 1972).

[20] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).

[21] P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).

[22] J. H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan, Ann Arbor, 1975).

[23] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, New York, 1989).

[24] S. Forrest, "Genetic algorithms: Principles of natural selection applied to computation," Science **261**, 872–878 (1993).

[25] D. E. Goldberg, *Computer-Aided Gas Pipeline Operation Using Genetic Algorithms and Machine Learning* (Ph.D. Dissertation in Civil Eng. University of Michigan, Ann Arbor, 1983).

[26] J. H. Holland, "Genetic algorithms," Sci. Am. **267**, 66–72 (1992).

[27] C. E. Floyd and G. D. Tourassi, "Computer-aided diagnosis using genetic algorithms and neural networks," in *Proceedings of the World Congress on Neural Networks*, Washington, DC (Lawrance Erlbaum Associates, NJ, 1995), pp. 863–866.

[28] F. Z. Brill, D. E. Brown, and W. N. Martin, "Fast genetic selection of features for neural network classifiers," IEEE Trans. Neural Networks **3**, 324–328 (1992).

[29] H. Fujita, T. Hara, X. Jing, T. Matsumoto, H. Yoshimura, and K. Seki, "Automated detection of lung nodules by using genetic algorithm technique in chest radiography," Radiology **197**, 426–426 (1995).

[30] D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Multiresolution texture analysis for classification of mass and normal breast tissue on digital mammograms," in Proceedings of SPIE Medical Imaging: Image Process. **2434**, (San Diego, CA, 1995), pp. 606–611.

[31] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsit, "Classification of mass and normal breast tissue: An artificial neural network with morphological features," in *Proceedings of the World Congress on Neural Networks*, Washington, DC (Lawrance Erlbaum Associates, NJ, 1995), pp. 876–879.

[32] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsit, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," IEEE Trans. Med. Imag. (in press).

[33] I. Daubechies, "Orthonormal bases of compactly supported wavelets," Commun. Pure Appl. Math. **41**, 909–996 (1988).

[34] Y. Hara, R. G. Atkins, S. H. Yueh, R. T. Shin, and J. A. Kong, "Application of neural networks to radar image classification," IEEE Trans. Geosci. Remote Sensing **32**, 100–109 (1994).

[35] N. Petrick, H.-P. Chan, B. Sahiner, D. Wei, M. A. Helvie, M. M. Goodsit, and D. D. Adler, "Automated detection of breast masses on digital mammograms using adaptive density-weighted contrast enhancement filtering," in Proc. SPIE Med. Imag. Image Process. **2434**, (San Diego, CA, 1995), 590–597.

[36] J. A. Freeman and D. M. Skapura, *Neural Networks: Algorithms, Applications and Programming Techniques* (Addison-Wesley, Reading, MA, 1991).

[37] M. M. Tatsuoka, *Multivariate Analysis, Techniques for Educational and Psychological Research* (Macmillan, New York, 1988).

[38] D. D. Dorfman and E. Alf, "Maximum likelihood estimation of parameters of signal detection theory—A direct solution," Psychometrika **33**, 117–124 (1968).

[39] C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binormal ROC curve from continuously distributed test results," presented at the 1990 Annual Meeting of the American Statistical Association, Anaheim, CA (1990).

[40] M. J. Norusis, *SPSS Professional Statistics 6.1* (SPSS Inc., Chicago, 1993).

[41] C. E. Metz, P. L. Wang, and H. B. Kronman, "A new approach for testing the significance of differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging: Proceedings of the 8th Conference*, edited by F. Deconinck (Martinus Nijhoff, Boston, 1984), pp. 432–445.

[42] H.-P. Chan, B. Sahiner, D. Wei, M. A. Helvie, D. D. Adler, and K. L. Lam, "Computer-aided diagnosis in mammography: Effect of feature classifiers on characterization of microcalcifications," Radiology **197**, 425–425 (1995).

[43] B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsit, "Classification of malignant and benign breast masses: Development of a high-sensitivity classifier using a genetic algorithm," accepted for presentation at the 82nd Annual Meeting of the Radiological Society of N. America, Chicago, IL (1996).

Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification

Nicholas Petrick, Heang-Ping Chan, Datong Wei, Berkman Sahiner, Mark A. Helvie, and Dorit D. Adler
*The University of Michigan, Department of Radiology, UH B1D403, 1500 E. Medical Center Drive,
Ann Arbor, Michigan 48109-0030*

(Received 5 July 1995; accepted for publication 17 May 1996)

This paper presents segmentation and classification results of an automated algorithm for the detection of breast masses on digitized mammograms. Potential mass regions were first identified using density-weighted contrast enhancement (DWCE) segmentation applied to single-view mammograms. Once the potential mass regions had been identified, multiresolution texture features extracted from wavelet coefficients were calculated, and linear discriminant analysis (LDA) was used to classify the regions as breast masses or normal tissue. In this article the overall detection results for two independent sets of 84 mammograms used alternately for training and test were evaluated by free-response receiver operating characteristics (FROC) analysis. The test results indicate that this new algorithm produced approximately 4.4 false positive per image at a true positive detection rate of 90% and 2.3 false positives per image at a true positive rate of 80%. © *1996 American Association of Physicists in Medicine.*

Key words: mammography, computer-aided diagnosis, density-weighted contrast enhancement filtering, linear discriminant analysis, texture classification

I. INTRODUCTION

Breast cancer is the most common malignancy affecting women and is second only to lung cancer in tumor related deaths in females. It was estimated that 182 000 new cases of breast cancer would occur in American women and 42 000 women would die from the disease in 1994.[1] This comprises 32% of all new cases of cancer and 18% of cancer deaths in women.[1] Efforts to decrease the mortality are currently aimed at early diagnosis and complete removal of small nonmetastatic lesions.[2] In an attempt to reduce cost and increase effectiveness, investigators are developing new techniques to improve detection of early breast cancers.[3] Computer-aided diagnosis (CAD) is one technique that may achieve both goals of lowering cost and increasing effectiveness.[4] CAD is especially well suited for the digital imaging technology which is being developed to produce digital images in full view mammography.

Several research groups have developed computer algorithms for automated detection of mammographic masses. Kegelmeyer has reported promising results for detecting spiculated lesions based on local edge characteristics and Laws texture features.[5,6] Both Lai *et al.*[7] and Qian *et al.*[8] proposed different variations of median filtering to enhance the digitized image prior to object identification. A thresholding method for mass localization and a mass classification algorithm using fuzzy pyramid linking have been developed by Brzakovic *et al.*[9] Other investigators have proposed using the asymmetry between the right and left breast images to determine possible mass locations. Yin *et al.* uses both linear and nonlinear bilateral subtraction[10] while the method by Lau *et al.* relies on "structural asymmetry" between the two breast images.[11] The above methods produced between one and five false detections for a true positive detection rate of approximately 90%. However, it is difficult to compare the effectiveness of these methods because each used a unique set of digitized mammograms, and the results varied between training and test. A general comparison between algorithms is further complicated by the fact that most of these studies were conducted using small data sets. While initial results from the first large scale preclinical study have been encouraging,[12] the performance of detection programs with clinical samples may not match their performance in laboratory tests.

Our preliminary study introduced the density-weighted contrast enhancement (DWCE) segmentation method and found that it was capable of detecting breast masses on 25 digitized mammograms.[13] In this article, a set of 168 digitized mammograms is used to evaluate a modified version of the original DWCE segmentation method in combination with a texture classification scheme. The following procedure was used to evaluate this new detection scheme. The set of digitized mammograms was first segmented into potential breast masses using the DWCE segmentation.[13] This method employed an adaptive filter to enhance structures within the breast region of a mammogram and then identified the structures using a simple edge detection algorithm. Once the digitized images were segmented using the DWCE, regions of interest (ROIs) based on the detected breast structures were extracted from each mammogram, and a set of multiresolution texture features were calculated for each extracted ROI. The feature set was then used by a linear discriminant analysis (LDA) algorithm to reduce the number of false detections. Finally, the performance of the DWCE segmentation and ROI texture classification scheme was evaluated using

0094-2405/96/23(10)/1685/12/$10.00

free-response receiver operating characteristics (FROC) analysis.

II. MATERIALS AND METHODS

A. Database

The clinical mammograms used in this study were acquired with American College of Radiology accredited mammography systems. Kodak MinR/MRE screen/film systems with extended cycle processing were used as the image recorder. The mammography systems have a 0.3-mm focal spot, a molybdenum anode, 0.03-mm-thick molybdenum filter, and a 5:1 reciprocating grid. The mammograms were selected from the files of patients who had undergone biopsy at the University of Michigan in the last five years. The selection criterion used by the radiologists was simply that a biopsy-proven mass existed on the mammogram. This set excluded lesions visible only by architectural distortions (i.e., no defined mass) but included masses accompanied by calcifications. No attempt was made to match the number of malignant and benign mass cases, but we did try to include a cross section of malignant masses. This led to a much larger proportion of malignant lesions than that in the general screening population. To avoid the effect of the repetitive grid pattern on the texture feature calculations, all mammograms with visible grid lines were excluded for the original set. Our final data set for this preliminary study was composed of 168 single-view mammograms. It included 85 malignant and 83 benign masses. The size of the masses ranged from 5 mm to 26 mm with a mean size of 12.2 mm, and their visibility ranged from 1 (obvious) to 10 (subtle) with a mean visibility of 4.51. A more complete discussion of the images selected for this study can be found in Wei *et al.*[14]

The mammograms were digitized with a LUMISYS DIS-1000 laser film scanner with a pixel size of 100 μm and 4096 gray levels. The digitizer logarithmically amplifies the light transmitted through the mammographic film before analog-to-digital conversion so that the gray levels are linearly proportional to optical densities in the range of 0.1 to 2.8 optical density units (O.D.). The O.D. range of the scanner is 0–3.5 with large pixel values in the digitized mammograms corresponding to low O.D. The digitized images used in this study were approximately 2000×2000 pixels in size. To conserve processing time and reduce noise in the initial DWCE segmentation stages, the full resolution mammograms were first smoothed with an 8×8 box filter and subsampled by a factor of 8, resulting in 800-μm images of approximately 256×256 pixels in size. However, the texture features used in the final LDA classification were calculated from the original images with a 100-μm pixel size.

The location and extent of all the biopsy-proven masses were marked on the original films by a radiologist. They were then localized on the digitized images and stored in a "truth" file on the computer by defining both the centroid (approximate center) of the lesion and the smallest bounding box (rectangle) containing the entire lesion. Both of these procedures were performed by hand using the original marked film as a guide. The centroid "truth" was used to analyze the initial DWCE segmentation. If an object segmented by the DWCE contained the centroid of the mass within the object region, it was considered a true positive (TP); otherwise, it was considered a false positive (FP). The centroid provided a fast method for evaluating the DWCE segmentation in its global and local stages. However, the final texture classification results are based on the more precise bounding box "truth." A region was considered a TP detection when at least 50% of the "truth" bounding box was detected. The centroid and bounding box definitions for the mass provided both an efficient mechanism for development of the DWCE and an accurate final analysis for the overall detection scheme.

For evaluation of the DWCE segmentation and subsequent texture classification, the 168 single-view mammograms were randomly divided into two groups of 84 images, groups G1 and G2, with the constraint that all images from a single patient were kept in the same group. A single set of DWCE segmentation parameters was applied to all images (G1 and G2) to extract potential mass regions. The regions extracted from the G1 and G2 images were then alternately used as training and test sets in the texture classification as described below.

B. Density-weighted contrast enhancement segmentation

Edge detection applied to an unenhanced image was not effective in detecting breast masses because of the low signal-to-noise ratio of the edges and the presence of complicated structured background. To overcome these problems, we have developed a new algorithm using DWCE filtering along with Laplacian–Gaussian (LG) edge detection for automatic segmentation of low contrast structures in digital mammograms.[13] The DWCE segmentation method employed adaptive filtering, edge detection, and morphological FP reduction to detect potential breast masses in a two-stage approach. In the first stage, DWCE segmentation was applied globally to the entire breast region of the mammogram to identify ROIs. In the second stage, the segmentation was applied locally to the ROIs identified in the global stage. Figures 1(a) and 1(b) depict the block diagrams for the global and local stages of this algorithm. The DWCE segmentation was originally introduced by Petrick *et al.*[13] but has been slightly modified in this study to improve its overall performance. In the following subsections we will summarize the main components of both the global and local stages, and highlight the differences between the original and current implementations of the DWCE technique.

1. Global stage: Density-weighted contrast enhancement filtering

The DWCE filter was developed to accentuate mammographic structures before edge detection by adaptively enhancing local contrast and is an extension of the local contrast and mean adaptive filter proposed by Peli and Lim.[15] The block diagram of the filter is shown in Fig. 2, while Fig. 3 contains examples of the images produced by each filter

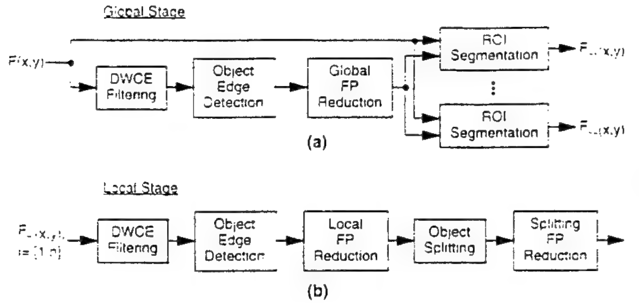


FIG. 1. The block diagram of the two-stage DWCE segmentation method used for the initial breast mass detection. The block diagram for the global stage is depicted in (a) while the local stage is shown in (b). Note, the outputs from the global stage, $F_L(x,y)$, are individually processed in the local stage.

block for a typical mammogram from our image set. All the DWCE functions introduced in the following discussion correspond to the steps illustrated in Fig. 2.

DWCE filtering was applied to the breast region [i.e., the breast map $F_{\mathrm{Map}}(x,y)$] of each mammogram which had been identified using thresholding and edge detection.[13] Figure 3(a) shows a typical mammogram, $F(x,y)$, at 800-μm resolution while 3(b) shows its breast map. The pixel intensities from $F(x,y)$ within the breast map were next rescaled to be between 0.0 and 1.0 producing a normalized breast image, $F_N(x,y)$. This normalization reduced the gray-level variation due to breast tissue composition and the imaging technique so that a single set of filter parameters could be applied uniformly to all digitized mammograms.

The normalized image was next split into a density and a contrast image, $F_D(x,y)$ and $F_C(x,y)$, respectively. $F_D(x,y)$ was produced by low-pass filtering the normalized input image using $G\{0,\sigma_D\}$, a Gaussian filter with zero mean and standard deviation $\sigma_D=8.0$. Likewise, $F_C(x,y)$ was produced by bandpass or high-pass filtering the normalized image. In the current DWCE implementation, $F_C(x,y)$ is created by subtracting the density image from the normalized input,

$$F_C(x,y)=F_N(x,y)-F_D(x,y), \tag{1}$$

or

$$F_C(x,y)=F_N(x,y)-G\{0,\sigma_D\}*F_N(x,y), \tag{2}$$

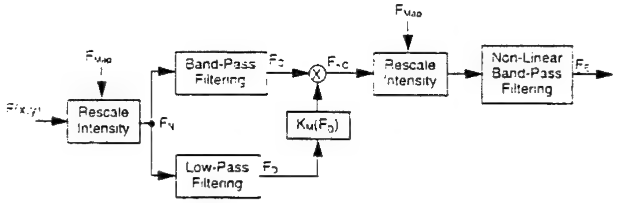


FIG. 2. The block diagram for the DWCE preprocessing filter used for image enhancement.

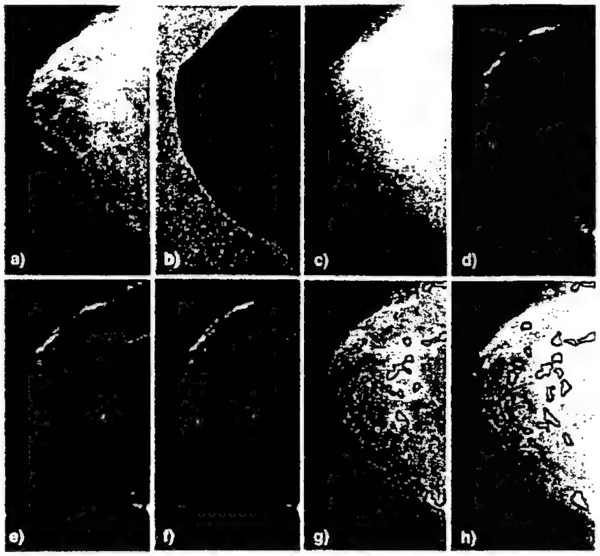


FIG. 3. (a) A typical mammogram from our image database; (b) the corresponding breast map used in the DWCE segmentation; (c) the density image ($F_D(x,y)$); (d) the contrast image ($F_C(x,y)$); (e) the weighted-contrast image ($F_{KC}(x,y)$); (f) the rescaled weighted-contrast image ($F_E(x,y)$); (g) the detected structures remaining after the global FP reduction step; (h) the detected structures remaining after the final splitting FP reduction step.

where $*$ represents two-dimensional convolution. Figures 3(c) and 3(d) show the density and contrast images obtained using this procedure.

The local density value, $F_D(x,y)$, was then used to determine a multiplication factor, $K_M(F_D(x,y))$, for each pixel (x,y) in the image. The multiplication factor was used to either enhance or suppress the local contrast and thereby produced a new weighted contrast image:

$$F_{\mathrm{KC}}(x,y)=K_M(F_D(x,y))\times F_C(x,y). \tag{3}$$

This process allowed the DWCE filter to adapt to local background characteristics within the image and was the principle component for our adaptive signal-to-noise ratio (SNR) enhancement. In this case, the signal refers to breast masses or other predominant structures within the breast. The output of the DWCE filter was given as

$$F_E(x,y)=K_{\mathrm{NL}}(F_{\mathrm{KC}}(x,y))\times F_{\mathrm{KC}}(x,y), \tag{4}$$

where each pixel, (x,y), in the weighted contrast image was used to define a second multiplication factor, $K_{\mathrm{NL}}(F_{\mathrm{KC}}(x,y))$, that nonlinearly scaled the weighted contrast image. This nonlinear scaling was used to further suppress the background and to separate merged structures in the DWCE enhanced image. Figures 3(e) and 3(f) show the weighted contrast and scaled weighted contrast images, respectively, obtained with the DWCE technique.

It can be seen that the two multiplication functions, K_M and K_{NL}, define the enhancement properties of the filter. These functions can be tailored to suit a specific task. Figures 4(a) and 4(b) show the curves selected for K_M and K_{NL}, respectively, in the current filter. The shape of the density-weighted contrast function, K_M, was selected to accentuate ($K_M(z)>1.0$, $z=F_D(x,y)$), the contrast at pixels in the den-

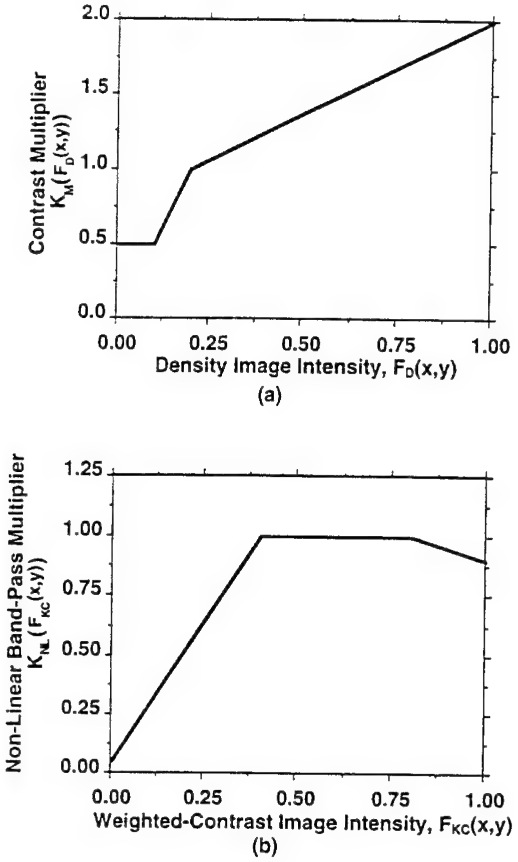


FIG. 4. Plots of (a) the weighted-contrast multiplication function ($K_M(z)$) and (b) the nonlinear rescaling function ($K_{NL}(z)$) used in the DWCE filtering.

sity image with medium to high intensity, while deemphasizing ($K_M(z)<1.0$), the contrast at pixels with low intensity. Thus, this function suppressed small structures mainly surrounded by background tissue and enhanced larger structures which are more likely to be masses. The exact shape of the multiplication function was determined experimentally by observing how detection was affected by variations in K_M. We chose $\{K_M(z)\geqslant 1.0 : 0.25\leqslant z\leqslant 1.0\}$ in the current weighted contrast function so that 75% of the intensity range was enhanced. K_M was found to be effective in reducing the background and enhancing breast structures, but it did not provide adequate separation between the structures. The shape of the nonlinear scaling function ($K_{NL}(z)$, $z=F_{KC}(x,y)$) was selected to provide additional separation between objects. Very low contrast regions were strongly deemphasized, thus eliminating many low-intensity bridges between individual structures. It was also found that a slight suppression of the highest contrast intensities provided a more uniform intensity distribution across detected breast structures. Again, the specific shape of the nonlinear contrast scaling was determined experimentally by observing the effect of different functional forms on the detection and object separation. A complete discussion of the DWCE multiplication functions used in this study can be found in the literature.[13]

2. Global stage: Object edge detection

The DWCE filtering was applied to the original mammogram to facilitate the detection of structures within the image and thus provided an estimate of their physical extent.[16] The DWCE implementation provided significant background reduction, as shown in Fig. 3(f), allowing for the use of a less complex edge detector. In this study, object edges were identified from the DWCE filtered mammogram using a Laplacian–Gaussian (LG) edge detector [Block 2 in Fig. 1(a)]. Edges in the enhanced image, $F_E(x,y)$, were defined as the zero crossing locations of

$$\nabla^2 G\{0,\sigma_E\} * F_E(x,y), \tag{5}$$

where $G\{0,\sigma_E\}$ was a zero mean Gaussian smoothing function with standard deviation $\sigma_E=2.0$.[17] The advantages of this edge detector are that its performance is independent of edge direction and that it tends to produce closed regions.

After the edge detection, all enclosed structures were filled to eliminate any holes that may have formed inside individual objects. The edges from each of the filled objects were tracked and identified. This edge detection is identical to the original DWCE implementation described in the literature.[13]

3. Global stage: False positive reduction

The DWCE filtering and subsequent edge detection do not differentiate between mass and normal tissues, therefore, a large number of potential regions were usually found. Since the shape of breast masses in general are different from those of normal tissue, we extracted morphological features and used a classification algorithm to identify some of these differences [Block 3 in Fig. 1(a)]. The goal here was to reduce the number of FP regions without losing a significant number of true masses, thus allowing the maximum number of TP regions to be passed on to the local processing stage. In this study, six additional morphological features were combined with the original set of five features used in the previous study[13] to improve the differentiation between mass and normal tissue objects. The original features were the number of edge pixels (P), the total object area ($A=\text{area}(F_{\text{obj}})$), the object's contrast, circularity, and rectangularity. The new features added in this implementation were the perimeter-to-area ratio (PAR) and a set of five normalized radial length (NRL) features. To define circularity and rectangularity, the minimum sized bounding box completely containing the object, $F_{\text{bb}}(x,y)$, and a circle with an area equivalent to the object area, $F_{\text{eq}}(x,y)$, were calculated. $F_{\text{eq}}(x,y)$ was centered at the object's centroid location and had radius r_{eq} given by

$$r_{\text{eq}}=\sqrt{\frac{\text{area}(F_{\text{obj}})}{\pi}}=\sqrt{\frac{A}{\pi}}. \tag{6}$$

Circularity and rectangularity were then defined as

$$\text{Circularity}\equiv\frac{\text{area}(F_{\text{obj}}\cap F_{\text{eq}})}{\text{area}(F_{\text{obj}})}, \tag{7}$$

$$\text{Rectangularity} \equiv \frac{\text{area}(F_{\text{obj}})}{\text{area}(F_{\text{bb}})}. \tag{8}$$

The five NRL features were a subset of the features defined by Kilday *et al.*[18] A radial length function was defined as the Euclidean distance from an object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object. This created an NRL vector given as

$$\mathbf{r} = \{r_k : 0 \leqslant k \leqslant N_E - 1\}, \tag{9}$$

where N_E was the number of edge pixels in the object. The histogram of the radial length was also calculated and created the probability vector

$$\mathbf{p} = \{p_j : 0 \leqslant j \leqslant N_H - 1\}, \tag{10}$$

where N_H was the number of bins used in the histogram. The NRL features selected in this study were the NRL mean value, standard deviation, entropy, area ratio, and zero crossing count. They are defined as

$$\mu_{\text{NRL}} \equiv \frac{1}{N_E} \sum_{k=0}^{N_E-1} r_k, \tag{11}$$

$$\sigma_{\text{NRL}} \equiv \sqrt{\frac{1}{N_E} \sum_{k=0}^{N_E-1} (r_k - \mu_{\text{NRL}})^2}, \tag{12}$$

$$E_{\text{NRL}} \equiv - \sum_{j=0}^{N_H-1} p_j \log(p_j), \tag{13}$$

$$AR_{\text{NRL}} \equiv \left\{ \frac{1}{N_E\, \mu_{\text{NRL}}} \sum_{k=0}^{N_E-1} (r_k - \mu_{\text{NRL}}) : r_k > \mu_{\text{NRL}} \right\}, \tag{14}$$

$$ZCC_{\text{NRL}} \equiv \text{number of zero crossings of } \{r_k - \mu_{\text{NRL}}\}_{k=0}^{N_E-1}. \tag{15}$$

A complete description of all the NRL features used in this study can be found in the literature.[18]

The extracted morphological features were used in a sequential classification scheme; a simple threshold classifier, followed by an LDA classifier, and finally followed by a backpropagation neural network (BPN). The purpose of each classifier was to reduce the number of FP regions with a minimum number of TP losses. This improved reduction scheme was selected because it has been found that sequential or parallel combinations of the different classifiers often increased the classification accuracy over the individual classifiers.[19,20] This is probably because they extract different information from the feature space. The threshold classifier simply set a maximum and a minimum value for each morphological feature. This provided some initial reduction and prevented the LDA and BPN classifiers from training with nonrepresentative object features. If all the morphological features from a detected object fell within the bounds, it was kept as a potential mass; otherwise, it was considered to be normal tissue and discarded. All DWCE detected objects with features values within the defined limits were saved as potential mass objects and passed on to the LDA classifier. The maximum and minimum feature limits, $f_{\text{Th}\uparrow}$ and $f_{\text{Th}\downarrow}$, respectively, were identical for both the G1 and G2 image groups and were selected as a multiple of the individual mass object bounds:

$$f_{\text{Th}\uparrow_i} = \{k \max_j (f_{i,j}) : \quad j \in [\text{index of all detected mass objects}]\}, \tag{16}$$

$$f_{\text{Th}\downarrow_i} = \{k \min_j (f_{i,j}) : \quad j \in [\text{index of all detected mass objects}]\}, \tag{17}$$

where $f_{i,j}$ is the value of the ith feature ($i \in [1,11]$) for the jth detected object. For this study, the multiplication factor (k) was selected to be 1.0. The second classifier, LDA, formed a linear combination of the morphological features and produced a single discriminant score for all remaining potential mass object. This classification scheme will be described in more detail in Sec. II C. The LDA classifier applied to the G1 objects was trained with the G2 object features and vice versa. This provided independent LDA training for each of the image sets. In order to minimize the probability of losing true masses, a lax discriminant threshold was chosen to retain most of the masses while achieving moderate FP reduction. The reduced sets of G1 and G2 objects with their morphological features were then passed on to a final BPN classification step. BPN formed a nonlinear combination of the morphological features into a single discriminant score. A complete description of the BPN morphological classification can be found in the literature.[13,21,22] In this step, a three input node, four hidden node, single output BPN architecture was utilized. The BPN classifier was trained in a similar fashion as the LDA but only the three most uncorrelated features (area, perimeter-to-area ratio, and contrast) were used as the input features. The individual G1 and G2 image sets were again used to train a pair of BPN classifiers, and the discriminant thresholds were chosen to maximize FP reduction while minimizing the loss of masses. All remaining DWCE detected objects after the application of the three classifiers were considered as potential mass objects and passed on to the ROI segmentation and subsequent local stage of the DWCE segmentation. Figure 3(g) shows the final reduced set of objects detected by the global stage for the original mammogram of Fig. 3(a).

4. Global stage: ROI segmentation

The final step in the global stage was the segmentation of the detected local regions [Blocks $4_{1\ldots,n}$ in Fig. 1(a)]. For each remaining potential mass object, a ROI corresponding to the object's bounding box was defined on the subsampled mammogram. The minimum size for these ROIs was chosen to be 32×32 pixels. A bounding box of an object smaller than this size was uniformly expanded in each direction (horizontal and vertical) until it reached 32×32 pixels. These defined object regions were then used as input ROIs to the local DWCE stage.

5. Local stage: DWCE filtering, edge detection, and local false positive reduction

The local stage of the DWCE segmentation was very similar to the global stage and is again depicted in Fig. 1(b). The main difference was that the processing was performed in local regions within the image. This local processing allowed the DWCE filter to adapt to the intensity distribution within each ROI and thus refined the borders of the detected objects. The input images to this stage, $F_{L_i}(x,y)$, were defined from the detected objects in the global stage. This local stage had five main components. Three of the components had corresponding global stage counterparts, and they included a second DWCE filter, LG edge detector, and local FP reduction step. The local DWCE filter and LG edge detector used identical parameters as their first stage counterparts, while the FP reduction step again used the 11 morphological features and the sequential thresholding, LDA, and BPN classification discussed previously. The only difference in the local FP reduction was that the feature and discriminant thresholds were adjusted to reflect the morphological properties of the locally extracted structures. Again, the goal of this FP reduction step was to reduce the number of potential mass regions before the regions were processed by a final texture classification stage. Therefore, lax decision thresholds were chosen to minimize additional losses of true mass objects.

6. Local stage: Object splitting and splitting FP reduction

The local processing of the mammograms lead to larger objects because of the improved estimate of the local background. However, the larger objects often resulted in region merging, (i.e., different structures within the breast merged into a single detected region). An object splitting step was therefore added to the local stage [Block 4 in Fig. 1(b)]. This splitting step enabled the use of fixed sized ROIs in the final texture classification. The splitting algorithm searched for narrowings in the cross section of an object. The algorithm initially found the cross-section width for each column in the object [$F_X(x)$ with length n]. Using $F_X(x)$, three parameters were calculated for each x. They were the area ratio of the two created objects along with the global and local cross-section width ratios. These ratios were defined as

$$F_{\mathrm{Area}}(x) \equiv \frac{\min(A_R(x),A_L(x))}{\max(A_R(x),A_L(x))}, \tag{18}$$

$$F_{\mathrm{Gbl}}(x) \equiv \left\{ 1.0 - \frac{F_X(x)}{\max(F_X(z))} : \; z \in [0,\, n-1] \right\}, \tag{19}$$

$$F_{\mathrm{Lcl}}(x) \equiv \left\{ 1.0 - \frac{F_X(x)}{\max(F_X(z))} : \; z \in [x-2,\, x+2] \right\}, \tag{20}$$

where $A_R(x)$ and $A_L(x)$ were the area of the right and left objects produced by splitting at location x. At each potential neck location, x, a cut value $F_{\mathrm{Cut}}(x)$ was defined as a linear combination of the cross-section ratios and the area ratio

$$F_{\mathrm{Cut}}(x) = 1.5F_{\mathrm{Gbl}}(x) + 2.0F_{\mathrm{Lcl}}(x) + 1.0F_{\mathrm{Area}}(x). \tag{21}$$

After similar cut functions were computed for each row and for the 45° and 135° directions, a maximum cut value was found for the object and compared to a cut threshold. If this maximum cut value exceeded this threshold, the object was split at that point; otherwise, it was left unchanged. If the object was split, the same algorithm was applied to the newly formed objects until no further splitting occurred. The splitting algorithm incorporated area information into the splitting process, thereby giving preference to narrowings closer to the center of the object and minimizing the number of times an object was split. For a complete description of this splitting algorithm refer to Petrick *et al.*[13]

The final FP reduction [Block 5 in Fig. 1(b)] again employed the 11 morphological features and the sequential classification scheme described in Sec. II B 3. The feature and discriminant thresholds were adjusted to reflect the morphological properties of the split objects. Figure 3(h) shows the set of detected objects after the complete two-stage DWCE segmentation for the original mammogram of Fig. 3(a).

C. Texture classification

After the DWCE segmentation identified a set of potential mass objects in the mammograms, ROIs corresponding to the detected object locations were extracted from the original 100-μm images and used as input to a texture classifier. The extracted ROIs had a fixed size of 256×256 pixels and the center of each ROI corresponded to the centroid location of a detected object. When the object was located close to the border of the mammogram and a complete 256×256 pixel ROI could not be defined, the ROI was shifted over until the appropriate edge coincided with the border of the original image. The classification of these fixed sized ROIs was based on a multiresolution texture analysis scheme. The approach has been described in detail in the literature[23] with the essential steps in the classification summarized below.

1. Texture features

The texture features used in the classification were derived from the spatial gray-level dependence (SGLD) matrix.[24,25] An element of the SGLD matrix, $p_{d,\theta}(i,j)$, is the joint probability that the gray levels i and j occur at a given interpixel separation d and direction θ. A set of SGLD matrices can be defined by varying the separation and direction. Thirteen texture features were derived from each SGLD matrix including correlation, energy, entropy, inertia, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy, and two measures of correlation information. The mathematical definitions for the SGLD features can be found in the literature.[23–26] These features were selected because they were found to be effective in the classification of ROIs containing masses or normal tissue manually identified by radiologists.[14,23,27,28] Each texture feature was calculated in the θ=0°,45°,90°,135° directions. The features obtained at θ=0°,90° and θ=45°,135° were averaged since no angular bias was seen in the texture of masses, and we did not find any significant difference in classification accuracy between features at separate angles and their averaged values.[28] The features calculated at adjacent pixels on axis (θ=0°,90°) and those in the diagonal direction (θ=45°,135°) were not averaged because of the significant $\sqrt{2}$ difference in the actual distances.[14]

Before the texture features were calculated, background correction was performed on the individual ROIs using a method described previously.[19,28] An ROI was first low-pass filtered and a pixel in the low-frequency background image was estimated as a weighted sum of the pixel values surrounding the ROI. The difference between the original ROI and the background thus reduced the gray-level variation due to the low-frequency structured background within the ROI.

2. Global multiresolution SGLD features

A wavelet transform with a four coefficient Daubechies kernel was used to decompose the individual ROIs into multiple scales after background correction.[14,29] Multiresolution ROI images were obtained using the original ROI (Scale 1) and the first two low-pass down-sampled approximation wavelets (Scales 2 and 4, respectively). The wavelet coefficients at Scale 8 were obtained by wavelet filtering but without down-sampling so that the minimum image size was maintained at 64×64 pixels. This minimum size was selected in order to reduce the statistical uncertainty when SGLD matrices of large pixel distances were calculated from the Scale 8 wavelet images.

Fourteen SGLD matrices, with effective distances of d={1,2,4,8,12,16,20,24,28,32,36,40,44,48} pixels relative to the original ROI, were calculated in both the on-axis (θ=0°,90°) and diagonal (θ=45°,135°) directions for each ROI using the Scale 1, 2, 4, and 8 wavelet images. Figure 5 contains a graphical representation of how the different wavelet images were related to the different SGLD matrices and the different object features. The SGLD matrices with d={1,2,4} were calculated using a pixel distance of one in the Scale 1, 2, and 4 wavelet images, respectively. The eleven SGLD matrices at d={8,12,16,20,24,28,32,36,40,44,48} were calculated from the Scale 8 wavelet image with pixel distances from 2 to 12 pixels. This process produced a total of 28 different SGLD matrices and 364 global multidistance texture features for each ROI.

3. Local multidistance SGLD features

A set of local texture features was also calculated for each ROI.[23,27] Five rectangular subregions were segmented from each ROI; an object subregion defined by the original DWCE object bounding box located at the center of the ROI, and four peripheral subregions at the corners. For a given pixel distance d and a given direction θ, an SGLD matrix was formed from the object subregion and another SGLD matrix was formed from the pixel pairs in the four peripheral subregions. These local SGLD matrices were calculated for d={1,2,4,8} and θ={0°,90°} and {45°,135°}. The thirteen texture features were calculated for both the object and periphery SGLD matrices. A total of 208 local features were defined for each ROI. They included the 104 features in the object region and 104 additional features defined as the difference between the feature values in the object and the periphery regions.

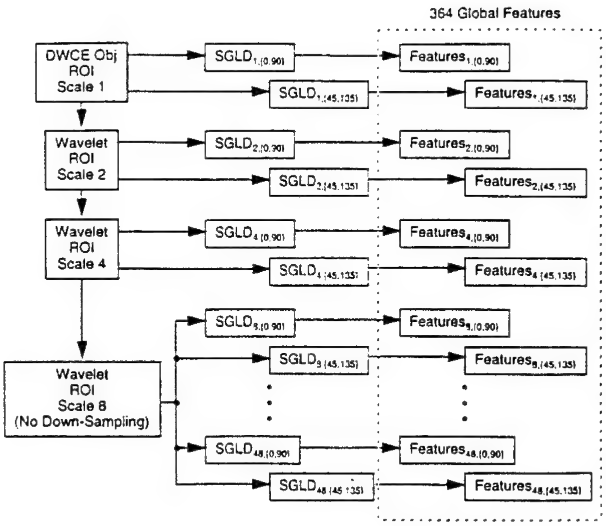


FIG. 5. Graphical representation of the parameters used in extracting features from the multiresolution wavelet images. The effective pixel distance d={1,2,4,8,12,16,20,24,28,32,36,40,44,48} for the SGLD matrices are relative to the original image.

4. Linear discriminant analysis

Linear discriminant analysis (LDA) uses a set of feature variables to classify an individual into one of a set of mutually exclusive classes.[30] We found in our previous studies that the LDA using SGLD texture features can effectively separate masses from normal tissue using ROIs manually selected by radiologists.[14,23] In our two class (mass and normal tissue) problem, the set of 572 global and local texture features was used as a pool of predictor variables in a stepwise selection procedure. This procedure selected a subset of features from the feature space based on the maximization of the Mahalanobis distance.[31] The stepwise selection eliminates irrelevant variables and thus improves the generalization capability of a discriminant function optimized with a finite number of training cases.

With the DWCE segmentation and object splitting algorithm, many of the extracted ROIs overlapped with one another because of the adjacency of the objects. We selected the independent ROIs (i.e., the ROIs that did not overlap with one another) to form a training set in order to avoid biases in the statistical distributions of the feature vectors. Two independent sets, $G1_i$ and $G2_i$, were formed by reducing all pairs of overlapping ROIs to single regions in G1 and G2, respectively. If a true mass ROI overlapped with a normal tissue ROI, the true mass region was saved while the normal region was eliminated. If two normal regions overlapped, one randomly selected region was eliminated. Finally, if two regions containing the full breast mass overlapped, the region defined by the DWCE segmented object which contained the centroid of the true mass was saved while the other was eliminated. These independent $G1_i$ and

TABLE I. The number of detected objects, the single stage reduction, the mean object area (μ_{Area}), and the standard deviation of the object areas (σ_{Area}) for the G1 data set after the global, local, and splitting stage FP reduction steps. The single stage reduction is defined as the reduction achieved by the morphological FP reduction block in each stage.

Stage	TP detections	FP detections per image	Single stage reduction	μ_{Area} (pixels)	σ_{Area} (pixels)
Global	82 of 84	34.6	25%	63.3	109.0
Local	81 of 84	12.4	75%	286.4	351.6
Split	81 of 84	18.9	14%	122.0	122.1

G2$_i$ sets were individually used to train the LDA classifiers while the full G1 and G2 sets were used for classifier evaluation.

To improve the statistical properties of the feature distributions, we used the entire set of segmented ROIs from both the G1$_i$ and G2$_i$ image sets for selection of feature variables. After feature selection, the G1 and G2 groups were used alternately as training and test sets. For example, when the coefficients of the linear discriminant function were optimized by the feature values from the G1$_i$ set, the classification accuracy of the linear discriminant function was tested with the full G2 set. The G1$_i$-trained linear discriminant function was also applied to the full G1 group to evaluate its self-consistency. Therefore, a total of four groups of discriminant scores were obtained: {Train:G1$_i$, Test:G1}, {Train:G2$_i$, Test:G2}, {Train:G2$_i$, Test:G1}, and {Train:G1$_i$, Test:G2}.

In this study, FROC analysis[32] was used to evaluate the performance of the complete segmentation method. The tradeoff between the TP fraction and the number of FP detections per image was determined by varying the decision threshold on the ROI discriminant scores. The raw detection data for both the full group training and test cases are presented, along with the fitted FROC curves obtained using the FROCFIT program.[32]

III. RESULTS

The number of TP and FP objects detected in the global and local stages of the DWCE segmentation are summarized in Tables I and II for the G1 and G2 image sets, respectively. A TP detection for the DWCE segmentation is again simply defined as an object locating the centroid of a breast mass, and a FP is any object other than the true mass (as discussed in Sec. II A). The two-stage DWCE segmentation missed only 8 of the 168 breast masses contained in the entire image set. Using the sets of TP and FP objects, 256×256 pixel ROIs representing each of the detected objects were extracted from the full resolution mammograms. A total of 1690 ROIs were extracted from the set of 84 G1 images and 1874 from the G2 mammograms. The independent G1$_i$ and G2$_i$ sets used for LDA training included 476 and 503 nonoverlapping ROIs, respectively. Stepwise feature selection was then performed on the 572 multidistance texture features using the combined G1$_i$ and G2$_i$ image sets, as described above, and 29 features were selected. These 29 features were used in the LDA texture classification for training and testing both the G1 and G2 image sets. Figures 6 and 7 show the raw and fitted training FROC curves obtained using the LDA texture classifier for the {Train:G1$_i$, Test:G1} and {Train:G2$_i$, Test:G2} combinations. The raw and fitted FROC curves for the test sets, {Train:G2$_i$, Test:G1} and {Train:G1$_i$, Test:G2}, are likewise depicted in Figs. 8 and 9. Finally, Table III contains the raw FROC results at TP detection rates of 90% and 80%, and Table IV contains the FROCFIT program parameters estimated for each of the fitted FROC curves.

TABLE II. The number of detected objects, the single stage reduction, the mean object area (μ_{Area}), and the standard deviation of the object areas (σ_{Area}) for the G2 data set after the global, local, and splitting stage FP reduction steps. The single stage reduction is defined as the reduction achieved by the morphological FP reduction block in each stage.

Stage	TP detections	FP detections per image	Single stage reduction	μ_{Area} (pixels)	σ_{Area} (pixels)
Global	79 of 84	32.9	32%	64.4	112.4
Local	79 of 84	21.4	62%	219.8	289.6
Split	79 of 84	21.6	7%	108.2	91.2

TABLE III. The number of FPs per image of each FROC curve at 90% and 80% TP detection fractions.

Training set	Test set	FPs per image (90% TP fraction)	FPs per image (80% TP fraction)
G1$_i$	G1	3.77	1.88
G2$_i$	G2	4.55	1.47
G2$_i$	G1	3.98	2.50
G1$_i$	G2	4.72	2.08

IV. DISCUSSION

A. DWCE segmentation

The purpose of the global processing stage was to define a set of local regions which contained the true breast masses and as few normal regions as possible. The initial DWCE filtering and subsequent edge detection was able to detect 161 of the 168 true masses in this preliminary study, including 83 of the 85 malignant masses. In addition, five of the seven missed masses, including both malignant masses, were

TABLE IV. Summary of the FROCFIT parameters and goodness of fit values. The headings for the table are; the two estimated FROCFIT parameters (a and b), the standard deviation of the estimated parameters (σ_a and σ_b), the area under the alternative FROC curve (A_{AFROC}), the standard deviation of the area (σ_A), the normalized chi-squared value (χ^2), and the significance probability for the fit (Prob).

Training set	Test set	a	σ_a	b	σ_b	A_{AFROC}	σ_A	χ^2	Prob
G1$_i$	G1	0.19	0.11	0.50	0.05	0.57	0.04	1.39	0.04
G2$_i$	G2	0.28	0.11	0.47	0.05	0.61	0.04	0.92	0.61
G2$_i$	G1	0.11	0.11	0.58	0.05	0.54	0.04	0.96	0.55
G1$_i$	G2	0.11	0.11	0.45	0.04	0.54	0.04	1.03	0.42

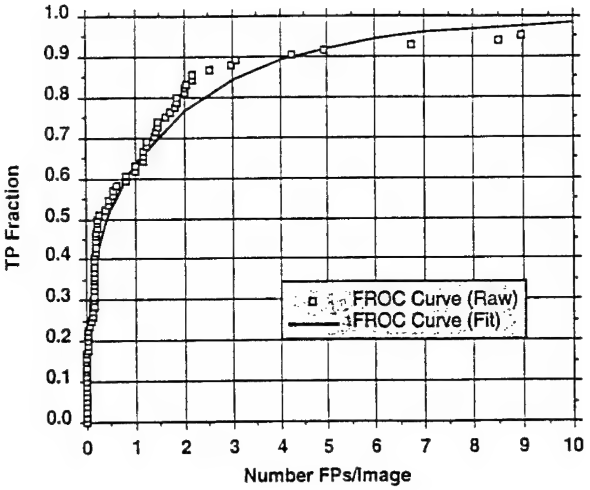


FIG. 6. FROC curves obtained with the image group {Train $G1_i$, Test G1}. The data points are raw data obtained by varying the decision threshold on the discriminant scores. The solid curve is obtained from the FROCFIT program.

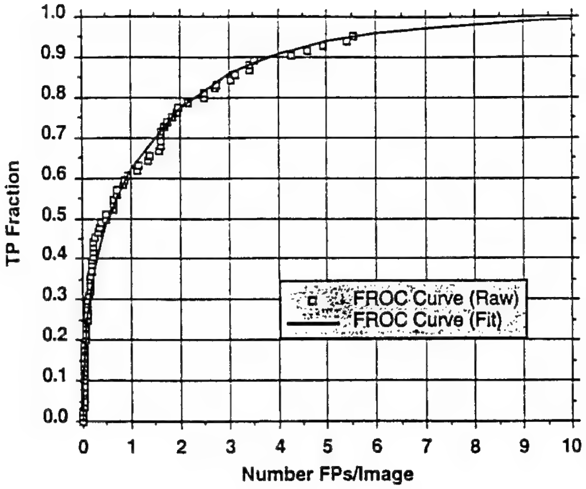


FIG. 8. FROC curves obtained with the image group {Train $G2_i$, Test G1}. The data points are raw data obtained by varying the decision threshold on the discriminant scores. The solid curve is obtained from the FROCFIT program.

detected in another mammogram containing a different view of the same breast. The image set did not include any additional views for the two remaining misses. This indicates that the global stage is effective in the initial detection task. However, the morphological properties of the detected regions proved to be of limited value in differentiating between TP and FP objects in the low-resolution DWCE filtered images. The main problem was that the global detection underestimated the size of the actual structures. This can be clearly seen in Fig. 3(g) where the detected objects are usually much smaller than the actual structures in the image. The average size, after FP reduction, of the global stage objects was 64.4 pixels. This underestimation can be mainly attributed to the large intensity range over which the background suppression was defined. This leads to inaccuracies in the object borders affecting the morphological features and reducing the effectiveness of the FP reduction. The morphological features and sequential classification were still able to achieve a 29% reduction in the initial number of regions, but at the end of the global stage an average of 34 detected regions per image across the G1 and G2 sets still remained. In further analysis of the detected regions, it was observed that fatty breasts had relatively few detected structures while mammograms containing dense tissue had a much larger number of regions.

The limitations of the global stage were partially overcome by repeating the filtering and edge detection in the local regions identified in the global stage. By allowing the DWCE filter to adapt to the background within these much

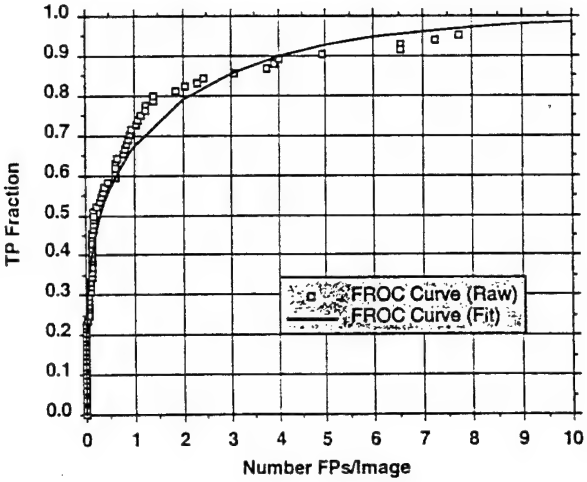


FIG. 7. FROC curves obtained with the image group {Train $G2_i$, Test G2}. The data points are raw data obtained by varying the decision threshold on the discriminant scores. The solid curve is obtained from the FROCFIT program.

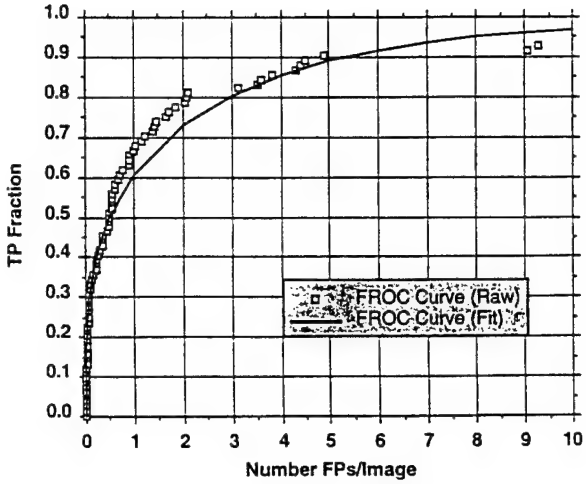


FIG. 9. FROC curves obtained with the image group {Train $G1_i$, Test G2}. The data points are raw data obtained by varying the decision threshold on the discriminant scores. The solid curve is obtained from the FROCFIT program.

smaller regions, better estimates for the true borders of the mammographic structures were achieved without sacrificing true mass detections. The local DWCE stage was able to detect 160 of the 168 true masses in this preliminary study where, again, 83 of the 85 malignant masses were detected. The additional missed mass did not come from the DWCE filtering and edge detection but was instead lost in the local FP reduction. This mass was detected in a different mammogram from our image set which contained a different view of the same breast. It is evident by comparing Figs. 3(g) and 3(h) that the detected objects in the local stage match the true borders better than the global stage objects. The average area of the detected objects following the local FP reduction increased to 253 pixels from the 64.4 pixels following the global stage. The more accurate borders help improve the local FP reduction which provided a 69% average reduction in the initial number of local FP regions and a corresponding 50% reduction in the number of FPs from the output of the global stage. The number of detected regions following the local FP reduction was still quite large, with an average of 16.9 regions detected per image. This large number of regions can be attributed to two factors. First, while improving the object border estimates, the local processing still continued to underestimate their true size [see Fig. 3(h)]. This limited the effectiveness of the morphological FP reduction in distinguishing between the masses and many of the normal structures. In addition, the expanded object area was attributed not only to the more precise edge characterization but also to the merging of neighboring regions into single detected objects. The merged objects caused problems in the final texture analysis stage because the texture information for the mass regions was often averaged with large amounts of normal tissue, thus increasing the likelihood that the true breast masses would be missed. Object splitting partially solved the problem of merged regions by estimating merge points according to geometrical shape. However, some distortion of the morphological features remained. Splitting also inadvertently introduced additional FPs. In this study the number of FPs increased from 16.9 FPs/image after local reduction to 20.3 FPs/image after the splitting reduction step. While the results of this preliminary study indicate that the DWCE segmentation is effective in detecting breast masses, further improvements in the scheme will be necessary to reduce the total number of detected regions.

Closer evaluation of the images where a mass was not detected highlighted a problem in the initial rescaling of some images. As stated previously, the initial rescaling step in the DWCE [refer to Fig. 1(a)] is very important because it allowed a single set of filters to be applied uniformly to all the mammograms. The rescaling should have occurred only within the breast region of the image. We have found that the initial breast map included a strip of pixels belonging to a bright edge outside the breast region of the mammogram in two of the images with missed masses. The pixels in this strip had a higher intensity than any of the other pixels in the breast region of the mammogram, and their inclusion in the rescaling caused many lower-intensity objects to be missed. When this strip was removed from the two images, the masses were detected.

B. Morphological classification

Another important factor that affects the FP reduction is the choice of morphological features. The eleven morphological features used in this study were selected because individually they showed some potential in differentiating between shapes. However, they are probably not the optimal set of morphological features for this task. The best subgroup of the features was found to be the area, perimeter-to-area ratio, and the contrast which provided the best BPN classifier performance. No general conclusions from this preliminary study can be made about the applicability of the individual features because of the small size of the image set and the suboptimal border information provided by the DWCE detection.

The morphological classification is an important component in the overall FP reduction. In this study, we selected the sequential application of a thresholding, an LDA, and a BPN classifier. The order of application was found to be important. The investigation showed that the LDA and especially the BPN classifier were trained faster and performed better when the initial number of FPs in the training set was small, thus leading to the use of the sequential classification scheme. We have not presented the exact values of the fixed thresholds used in this study because of the small size of this preliminary image set. With a larger, more representative training set, the particular threshold values will need to be adjusted. Therefore, we have instead concentrated on describing the general methodology for selecting the individual thresholds, as outlined in Sec. II B.

C. Texture classification

The large number of regions detected in the DWCE segmentation precipitated the need for additional FP reduction. This additional reduction was achieved by classifying with multiresolution texture features extracted from the DWCE detected regions. The LDA classification using SGLD features was selected because it was found to be effective in differentiating breast masses from normal tissue in regions identified by radiologists.[14,23] Again, we have not presented details about the particular feature selected because of the small size of the data set. However, a detailed discussion of the multiresolution texture features and the LDA texture classification method can be found in the literature.[14,23] The texture classification in this final step resulted in an average of 4.4 FPs/image at a 90% TP rate and 2.3 FPs/image at an 80% TP rate (Table III) in the test sets. These results indicate that the overall system (i.e., DWCE segmentation plus LDA texture classification) is capable of automatically detecting breast masses on digitized mammograms. Table III also indicates that the G1 and G2 image sets were reasonably well matched. The G2 set provided slightly better performance at a 90% TP fraction but the G1 set's performance was better for the 80% detection level.

Figures 6–9 contain fitted FROC curves obtained using the FROCFIT program developed by Chakraborty *et al.*[32] Table IV contains the estimated fit parameters and the goodness of fit characteristics obtained with the program. The fitted curves match well visually with the raw FROC results and the normalized χ^2 goodness of fits only varied from 0.92 to 1.39 (optimal value is 1.0). This indicates that the FROCFIT program may be able to fit our raw FROC results. However, it is likely the signals detected by our method do not satisfy the assumptions that the occurrence of an FP follows Poisson statistics and that the FPs are independent. Further studies are therefore needed to investigate if the good fit observed occurs by chance and if the area under the alternative FROC curve (A_{AFROC}) can be used as an indication of the overall performance of the classification system.

D. Future studies

Our results indicate that DWCE segmentation can be used to effectively detect breast structures on a mammogram. The flexible form of the DWCE filter leaves open the possibility that further optimization of the detection parameters may improve overall performance. Evaluation of different DWCE filters (e.g., modifying K_M and K_{NL}) will be pursued in future studies.

One of the difficulties in the DWCE segmentation method is the merging of regions and the subsequent need to split objects. The splitting operation increases the number of false regions and also adversely affects the morphological information by introducing straight edges at the split locations. In future studies, we will investigate alternative methods for separating merged structures. Gray-level information will be used in conjunction with binary shape information to guide the splitting. The sequential change in shape obtained by region growing at different local threshold levels will more precisely define multiple regions within a single DWCE segmented object. This approach should improve the morphological features of the split objects and increase the classification accuracy of masses and normal tissue, thereby reducing the FP detections. Furthermore, a fundamental improvement in the adaptivity of the DWCE segmentation will be needed to reduce the number of objects extracted in the initial stage. One possible improvement may be accomplished by first classifying the breast parenchyma into different types (e.g., fatty, mixed, or dense). The DWCE filter parameters can then be optimized specifically for each tissue type. This would allow better background suppression and more precise object extraction in different types of breast parenchyma. It can be expected that the initial number of FP objects detected in dense breasts will be reduced without impacting the detection on fatty breasts.

Our detection scheme makes use of information on a single mammogram. In mammographic interpretation, it has been found that symmetry information on the left and right mammograms of the same view often improves the detection of subtle abnormal tissue density.[33] The information can also be used to eliminate FP detections when they appear on both mammograms in symmetrical locations.[10] However, the symmetry information should be used with caution because many patient mammograms are not highly symmetrical due to variations in compression and imaging techniques, as well as the natural asymmetry in tissue structures. We will investigate the effectiveness of the symmetry information from paired mammograms in FP reductions in future studies.

V. CONCLUSION

We have developed an image enhancement technique which can adaptively suppress the low-frequency structured background and enhance the contrast of structures on an image. The technique was applied to the segmentation step in a CAD program for detection of breast masses. It was found to be effective in enhancing masses and normal tissue structures on mammograms. To further distinguish between masses and normal tissue, the potential mass regions were classified with an LDA using multiresolution texture features extracted from wavelet coefficients at several scales. Results of FROC analysis indicate that the current algorithm can achieve a TP rate of 90% at 4.4 FPs/image and a TP rate of 80% at 2.3 FPs/image. The consistency in the performance of the algorithm was verified by training and testing two independent data sets. This study demonstrates the feasibility of our approach to computer-assisted detection of masses in mammographic interpretation. Further investigations are under way to improve the detection accuracy and test its performance in large data sets.

ACKNOWLEDGMENTS

This work is supported by a USPHS Grant No. CA 48129 and U.S. Army Grant No. DAMD 17-93-J-007 (through subgrant No. GU RX 4300-803UM from Georgetown University). The content of this publication does not necessarily reflect the position of Georgetown University or the government, and no official endorsement of any equipment or product of any company mentioned in the publication should be inferred. The authors thank Dev Chakraborty, Ph.D. for providing the FROCFIT program.

[1] C. C. Boring, T. S. Squires, T. Tong, and S. Montgomery, "Cancer statistics, 1994," CA: Cancer J. Clin. **44**, 7–26 (1994).

[2] B. A. Porter, V. Taylor, J. P. Smith, and V. Tsao, "Contrast-enhanced magnetic resonance mammography," Acad. Radiol. **1S1**, S36–S46 (1994).

[3] F. Shtern, C. Stelling, B. Goldberg, and R. Hawkins, "Novel technologies in breast imaging: National cancer institute perspective," in *2nd Post-Graduate Course Syllabus* (Society of Breast Imaging, Orlando, FL, 1995), pp. 153–156.

[4] C. J. Vyborny and M. L. Giger, "Computer vision and artificial intelligence in mammography," Am. J. Roentgenol. **162**, 699–708 (1994).

[5] W. P. Kegelmeyer, Jr., "Computer detection of stellate lesions in mammograms," Proc. SPIE Biomed. Image Process. **1660**, 446–454 (1992).

[6] W. P. Kegelmeyer, Jr., J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).

[7] S. M. Lai, X. Li, and W. F. Bischof, "On techniques for detecting circumscribed masses in mammograms," IEEE Trans. Med. Imag. **8**, 377–386 (1989).

[8] W. Qian, L. P. Clarke, M. Kallergi, and R. A. Clark, "Tree-structured nonlinear filters in digital mammography," IEEE Trans. Med. Imag. **13**, 25–36 (1994).

[9] D. Brzakovic. X. M. Luo, and P. Brzakovic. ''An approach to automated detection of tumors in mammography,'' IEEE Trans. Med. Imag. **9**, 233–241 (1990).

[10] F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, R. A. Vybomy, and C. J. Schmidt. ''Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images,'' Med. Phys. **18**, 955–963 (1991).

[11] T.-K. Lau and W. F. Bischof, ''Automated detection of breast tumors using the asymmetry approach,'' Comput. Biomed. Res. **24**, 273–295 (1991).

[12] R. M. Nishikawa, R. C. Haldemann, J. Papaioannou, M. L. Giger, P. Lu, D. E. Woverton, R. A. Schmidt, U. Bick, K. J. Munn, and K. Doi, ''Initial experience with a prototype clinical intelligent mammography workstation for computer-aided diagnosis,'' Proc. SPIE Med. Imag. Image Process. **2434**, 65–71 (1995).

[13] N. Petrick, H.-P. Chan, B. Sahiner, and D. Wei. ''An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection,'' IEEE Trans. Med. Imag. **15**, 59–67 (1996).

[14] D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt. ''Classification of mass and normal breast tissue on digital mammograms: Multi-resolution texture analysis,'' Med. Phys. **22**, 1501–1513 (1995).

[15] T. Peli and J. S. Lim, ''Adaptive filtering for image enhancement,'' Opt. Eng. **21**, 108–112 (1982).

[16] W. K. Pratt. *Digital Image Processing* (Wiley, New York, 1978).

[17] D. Marr and E. Hildreth, ''Theory of edge detection,'' Proc. R. Soc. London Ser. B Biolog. Sci. **207**, 187–217 (1980).

[18] J. Kilday, F. Palmieri, and M. D. Fox, ''Classifying mammographic lesions using computerized image analysis,'' IEEE Trans. Med. Imag. **12**, 664–669 (1993).

[19] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt. ''Classification of mass and normal breast tissue on digital mammograms: A convolution neural network classifier with spatial domain and texture images,'' IEEE Trans. Med. Imag. (in press).

[20] L. Xu, A. Krzyżak, and C. Y. Suen, ''Methods of combining multiple classifiers and their applications to handwriting recognition,'' IEEE Trans. Sys. Man Cybern. **22**, 418–435 (1992).

[21] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt. ''Image classification using artificial neural networks,'' Proc. SPIE Med. Imag. Image Process. **2434**, 838–845 (1995).

[22] J. A. Freeman and D. M. Skapura, *Neural Networks: Algorithms, Applications and Programming Techniques* (Addison-Wesley, Reading, MA, 1991).

[23] D. Wei, H.-P. Chan, N. Petrick, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt. ''False-positive reduction techniques for the detection of masses on digital mammograms: Global and local multi-resolution texture analysis,'' Med. Phys. (submitted).

[24] A. Petrosian, H.-P. Chan, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, ''Computer-aided diagnosis in mammography: Classification of mass and normal tissue by texture analysis,'' Phys. Med. Biol. **39**, 2273–2288 (1994).

[25] R. W. Conner, ''Toward a set of statistical features which measure visually perceivable qualities of textures,'' in Proc. IEEE Conf. Pattern Recogn. Image Process. 382–390 (1979).

[26] R. M. Haralick, K. Shanmugam, and I. Dinstein, ''Texture features for image classification,'' IEEE Trans. Sys. Man Cybern. **3**, 610–621 (1973).

[27] D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, ''Multiresolution texture analysis for classification of mass and normal breast tissue on digital mammograms,'' Proc. SPIE Med. Imag. Image Process. **2434**, 606–611 (1995).

[28] H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, ''Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space,'' Phys. Med. Biol. **40**, 857–876 (1995).

[29] I. Daubechies, ''The wavelet transform, time-frequency localization, and signal analysis,'' IEEE Trans. Inf. Theory **36**, 961–1005 (1990).

[30] P. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).

[31] M. J. Norušis, *SPSS Professional Statistics 6.1* (SPSS Inc., Chicago, IL, 1993).

[32] D. P. Chakraborty, ''Maximum likelihood analysis of free-response receiver operating characteristics (froc) data,'' Med. Phys. **16**, 561–568 (1989).

[33] L. Tabár and P. B. Dean, *Teaching Atlas of Mammography*, 2nd ed. (Georg Thieme Verlag, New York, 1985).

Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network

Heang-Ping Chan, Shih-Chung B. Lo,[a] Berkman Sahiner,
Kwok Leung Lam,[b] and Mark A. Helvie
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 30 December 1994; accepted for publication 14 June 1995)

We are developing a computer program for automated detection of clustered microcalcifications on mammograms. In this study, we investigated the effectiveness of a signal classifier based on a convolution neural network (CNN) approach for improvement of the accuracy of the detection program. Fifty-two mammograms with clustered microcalcifications were selected from patient files. The clusters on the mammograms were ranked by experienced mammographers and divided into an obvious group, an average group, and a subtle group. The average and subtle groups were combined and randomly divided into two sets, each of which was used as training or test set alternately. The obvious group served as an additional independent test set. Regions of interest (ROIs) containing potential individual microcalcifications were first located on each mammogram by the automated detection program. The ROIs from one set of the mammograms were used to train CNNs of different configurations with a back-propagation method. The generalization capability of the trained CNNs was then examined by their accuracy of classifying the ROIs from the other set and from the obvious group. The classification accuracy of the CNNs for the ROIs was evaluated by receiver operating characteristic (ROC) analysis. It was found that CNNs of many different configurations can reach approximately the same performance level, with the area under the ROC curve (A_z) of 0.9. We incorporated a trained CNN into the detection program and evaluated the improvement of the detection accuracy by the CNN using free response ROC analysis. Our results indicated that, over a wide range of true-positive (TP) cluster detection rate, the CNN classifier could reduce the number of false-positive (FP) clusters per image by more than 70%. For the obvious cases, at a TP rate of 100%, the FP rate reduced from 0.35 cluster per image to 0.1 cluster per image. For the average and subtle cases, the detection accuracy improved from a TP rate of 87% at an FP rate of four clusters per image to a TP rate of 90% at an FP rate of 1.5 clusters per image.

Key words: mammography, microcalcification, computer-aided diagnosis, artificial neural network, receiver operating characteristic (ROC) analysis

I. INTRODUCTION

In the United States, breast cancer is the leading cause of death in women between 40 and 55 yr of age.[1] One out of eight women will develop breast cancer in their lifetime.[2] Studies have indicated that early detection and treatment improve the chances of survival for breast cancer patients. At present, mammography is the only proven method that can detect minimal breast cancers.[3–5] However, 10%–30% of the breast cancers that are visible on mammograms in retrospective studies are not detected due to various technical or human factors.[6–9] Double reading can reduce the miss rate on radiographic reading.[10] It has also been shown that computer-aided diagnosis (CAD), in which a computer alerts radiologists to suspicious locations on the images during mammographic reading, can improve the detection accuracy significantly.[11,12] CAD is thus a viable cost-effective alternative to double reading by radiologists.

One of the important indicators of the presence of breast cancers is clustered microcalcifications.[13] Clustered microcalcifications can be seen on mammograms in 30%–50% of breast cancers.[14–17] It is difficult to detect subtle microcalcifications because of the noisy mammographic background. A number of research groups have been developing CAD programs for the detection of microcalcifications. Chan *et al.*[11,18,19] demonstrated that a difference-image technique can effectively detect microcalcifications on digitized mammograms. Fam *et al.*[20] and Davies *et al.*[21] detected microcalcifications using conventional image processing techniques. Qian *et al.*[22] recently devised a tree-structure filter and wavelet transform for enhancement of microcalcifications to facilitate detection. Other groups extracted morphological features such as contrast, size, shape, and edge gradient of microcalcifications, and classified them with various feature classifiers.[23–31] Wu *et al.* scanned for suspected microcalcifications with the difference-image technique[18] then further classified true and false detections by an artificial neural network based on features extracted from their power spectra.[32] Similarly, Zhang *et al.*[33] used a shift-invariant neural network to reduce false-positive microcalcifications. The results reported in all these studies appear to be encouraging for the selected datasets.

In this study, we trained a convolution neural network (CNN) to recognize mammographic microcalcifications. The CNN was first developed for the detection of pulmonary nodules on chest radiographs.[34] This neural network is different from the commonly used back-propagation neural network in that its input is a region of interest (ROI) from the image instead of extracted image features. It is also different from the shift-invariant neural network used by Zhang *et al.*[33] in that the input ROI to the CNN includes an individual microcalcification instead of a cluster, and that the

0094-2405/95/22(10)/1555/13/$6.00

output of the CNN is a decision score for determination of the presence of a microcalcification instead of a processed image ROI. Therefore, with our approach no further image processing techniques such as thresholding and region growing have to be applied to an output ROI to determine if a microcalcification is present. We have incorporated the trained CNN into our detection program and its effectiveness is evaluated by the improvement in the overall detection accuracy of the CAD program.

II. MATERIALS AND METHODS

A. Case selection

In this study, we used mammograms that contained clustered microcalcifications as case samples. The mammograms were selected from the patient files in the Department of Radiology at the University of Michigan Hospitals by experienced mammographers. The mammograms were acquired with a dedicated mammographic system with a 0.3 mm focal spot, molybdenum (Mo) anode and 0.03 mm Mo filter, and a 5:1 reciprocating grid. Kodak Min R/MRE mammographic screen/film system using extended cycle processing was employed as the image receptor. The presence of the clustered microcalcifications and the histology for each case had been verified by biopsy. The case samples included a mixture of benign and malignant cases. However, in this study, we concentrated on the detection rather than the classification of the malignant/benign nature of the microcalcifications.

Fifty-two mammograms were selected for this study. Each mammogram was ranked by the radiologist regarding the visibility of the cluster of microcalcifications on a rating scale of 1–5 (1=very obvious, 5=very subtle). The scale was established subjectively relative to the cases encountered in clinical practice in our hospitals. After ranking, we divided the 52 mammograms into three groups: the mammograms of ratings 1 and 2 were referred to as the obvious group ($N=14$), the mammograms of rating 3 as the average group ($N=16$), and the mammograms of ratings 4 and 5 as the subtle group ($N=22$). Although this classification was very subjective, it was an attempt to demonstrate the dependence of the performance of the CAD program on the database. We also attempted to describe quantitatively the physical characteristics of the microcalcifications on the digitized image and correlated them with the visual ratings. We extracted digitally, as discussed below, the contrast, the size, and the signal-to-noise ratio (SNR) of the individual microcalcifications. The mean and standard deviation (SD) of these physical characteristics of the microcalcifications in each group were compared.

B. Digitization of mammograms

All mammograms were digitized with a laser film scanner (LUMISYS DIS-1000), with both the sampling distance and the nominal spot size, and thus the pixel size, chosen to be 0.1 mm×0.1 mm.[35] The digitizer has a gray level resolution of 12 bits and an optical density (O.D.) range of 0–3.5. It was calibrated so that the O.D. on film was linearly proportional to output pixel values in the range of about 0.1 O.D. to 2.8 O.D. at 0.001 O.D./pixel value. The slope of the calibration curve outside this range decreased gradually. Before input to the detection program, the pixel values were linearly converted. such that low optical density was represented by high pixel values.

To establish a "truth" file with which the computer detection results could be compared, we determined the true locations of the individual microcalcifications on each mammogram manually. The digitized image was displayed on a workstation and the region containing the cluster of microcalcifications was enlarged to full resolution. Each individual microcalcification on the displayed image was identified carefully by comparison with the mammogram on film with a magnifier. The coordinates of the microcalcifications were then determined by a cursor and stored in the "truth" file. The same regional clustering procedure as that used in the detection program described below was applied to the "truth" file to determine the coordinates of the centroid of the clusters. These coordinates were used for scoring the detection of the clusters by the automated procedure.

It may be noted that the "truth" file thus determined may not be the absolute truth because of the difficulties and uncertainties in detecting subtle microcalcifications that are near the human visual threshold. However, this is the best available and practical method. Neither histologic analysis nor specimen radiographs can be used to identify individual microcalcifications seen on mammograms because of the very different geometry and image quality obtained with these techniques. Magnification mammograms are often not available since magnification is not performed for every case or for all views.

C. Extraction of signal characteristics

To describe quantitatively the physical characteristics of the microcalcifications on the digitized image, we have developed a signal extraction program to determine the size, contrast. SNR of the microcalcifications from an unprocessed image based on the coordinate of each individual microcalcification in the "truth" file.[36] In a 51×51 pixel ROI centered at each signal site. the structured background is estimated by polynomial curve fitting in the x and y directions. The fitted pixel values in the x and y directions at the same pixel are averaged. The process may be performed more than one time to reach a well-fitted smooth surface. The central $l\times l$ pixels in the region which contain the signal are excluded from the curve fitting and noise estimation. The size l is chosen to be a constant that is larger than the diameters of the microcalcifications of interest yet much smaller than 51 pixels. After subtraction of the structured background, the local root-mean-square (RMS) noise is calculated. A local threshold gray level is determined as the product of the RMS noise and an input SNR threshold. With a region growing technique. the signal region is then extracted as the connected pixels above the threshold around the manually identified signal location. The size of the microcalcification is estimated as the number of pixels in the signal region. The contrast is defined as the maximum pixel value in the signal region after subtracting the background. The SNR of the microcalcification is the ratio of the contrast to the local RMS noise. The thresholded image of the microcalcifications superimposed on a background of constant pixel values can also be displayed for visual comparison.

D. Computerized detection of microcalcifications

We have developed a computer program that can automatically detect microcalcifications on mammograms. The program has been described in the literature.[11,18,35] Briefly, there are three major steps in the algorithm: preprocessing, segmentation, and classification. In the preprocessing step, an edge detector detects the breast boundary and divides the image into two regions, one internal and the other external to the breast. Signal detection is applied only to the region within the breast. A signal-enhancement filter (1×1 kernel) is employed to enhance the microcalcifications and a signal-suppression filter (box-rim filter with an 8×8 kernel of constant weights around the rim and a 4×4 central area of zero weights), to remove or suppress the microcalcifications and smooth the noise. Subtracting the two filtered images results in an SNR-enhanced image in which the low-frequency structured background is removed and the high-frequency noise is suppressed. This is also referred to as a difference-image technique.[11,18,19,35] When both the signal-enhancement filter and the signal-suppression filter are linear, as used in this study, the difference-image technique is equivalent to bandpass filtering. In the segmentation step, the program determines the gray level histogram of the preprocessed image within the breast region. A gray level thresholding technique is used to locate potential signal sites above a global threshold. The threshold is changed iteratively until the number of sites obtained falls within the chosen input maximum (4000) and minimum (3000) numbers. At each potential site, a locally adaptive gray level thresholding technique in combination with region growing is performed to determine the number of connected pixels above a local threshold, which is calculated as the product of the local RMS noise and an input SNR threshold. The signal characteristics to be used in the classification step, such as the size, maximum contrast, SNR, and its location, are obtained in this step. This locally adaptive thresholding technique is similar to the signal characteristic extraction technique described above, except that the procedure is performed on the SNR-enhanced image instead of the unprocessed image so that no curve fitting for background correction is necessary.

In the classification step, the previous computer program performs three tests to distinguish signals from noise or artifacts. A lower bound (two pixels) is imposed on the size to exclude signals below a certain size that are likely to be noise and an upper bound (80 pixels) is set to exclude signals greater than a certain size that are likely to be large benign calcifications. A contrast upper bound is also set to exclude potential signals that have a contrast higher than an input number (10) of SDs above the average contrast of all potential signals found with local thresholding. This criterion excludes the very high-contrast signals that are likely to be artifacts and large benign calcifications. A regional clustering procedure is then applied to the remaining signals; a signal is kept if the number of signals found within a neighborhood of a chosen input diameter (1 cm) around that signal is greater than an input minimum number. The remaining signals that are not found to be in the neighborhood of any potential clusters will be considered isolated noise points or calcifications and excluded. This clustering criterion is useful for reducing false positives, because true microcalcifications of clinical interest always appear in clusters on mammograms.[13–17] The specific parameters used in each step have been described previously.[11,19,35]

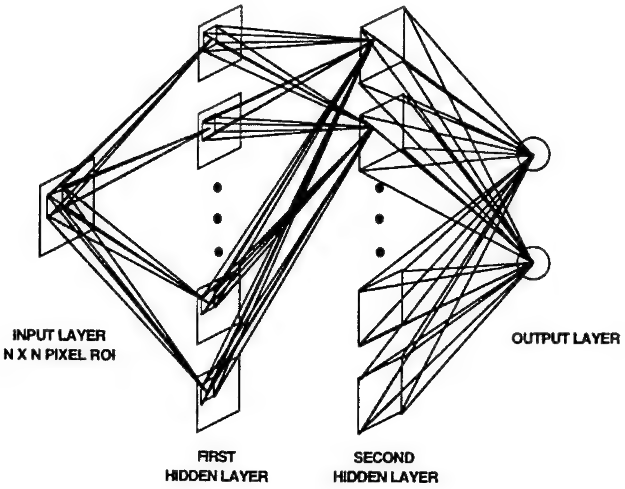


FIG. 1. Schematic diagram of the architecture of a convolution neural network (CNN). The input ROI size, the number of hidden layers, and the number of node groups in each layer are varied in this study.

In this study, we investigated the effectiveness of a trained convolution neural network (CNN)[34] in discriminating false signals from true microcalcifications. The chosen CNN classifier was incorporated in the detection program. The potential signals that passed the size and contrast tests in the classification step were further screened by the CNN before being examined by the regional clustering criterion. The overall detection accuracy of microcalcifications with and without the CNN classifier could then be compared.

E. Convolution neural network classifier

The artificial neural network (ANN) used in this application is a convolution-type neural network.[34] The CNN can be considered a simplified version of the neocognitron[37] designed to simulate the human visual system. The general architecture of the CNN used in this study is shown in Fig. 1. It consists of an input layer, one to several hidden layers, and an output layer. The input layer of the CNN contains $N\times N$ input nodes, each of the input nodes is a sensor for an input pixel value in an $N\times N$-pixel ROI containing the normal or abnormal pattern to be recognized. In the hidden layers, the nodes are organized in groups and the groups between adjacent layers are interconnected by weights that are organized in kernels. Learning is constrained such that the kernel of weights connecting the kth group in the $(L-1)$th layer to the nth group in the Lth layer is invariant with nodes in the same groups. Forward signal propagation is thus similar to a spatially invariant convolution operation; the signals from the nodes in the lower layer are convolved with the weight kernel, and the resultant value of the convolution is collected

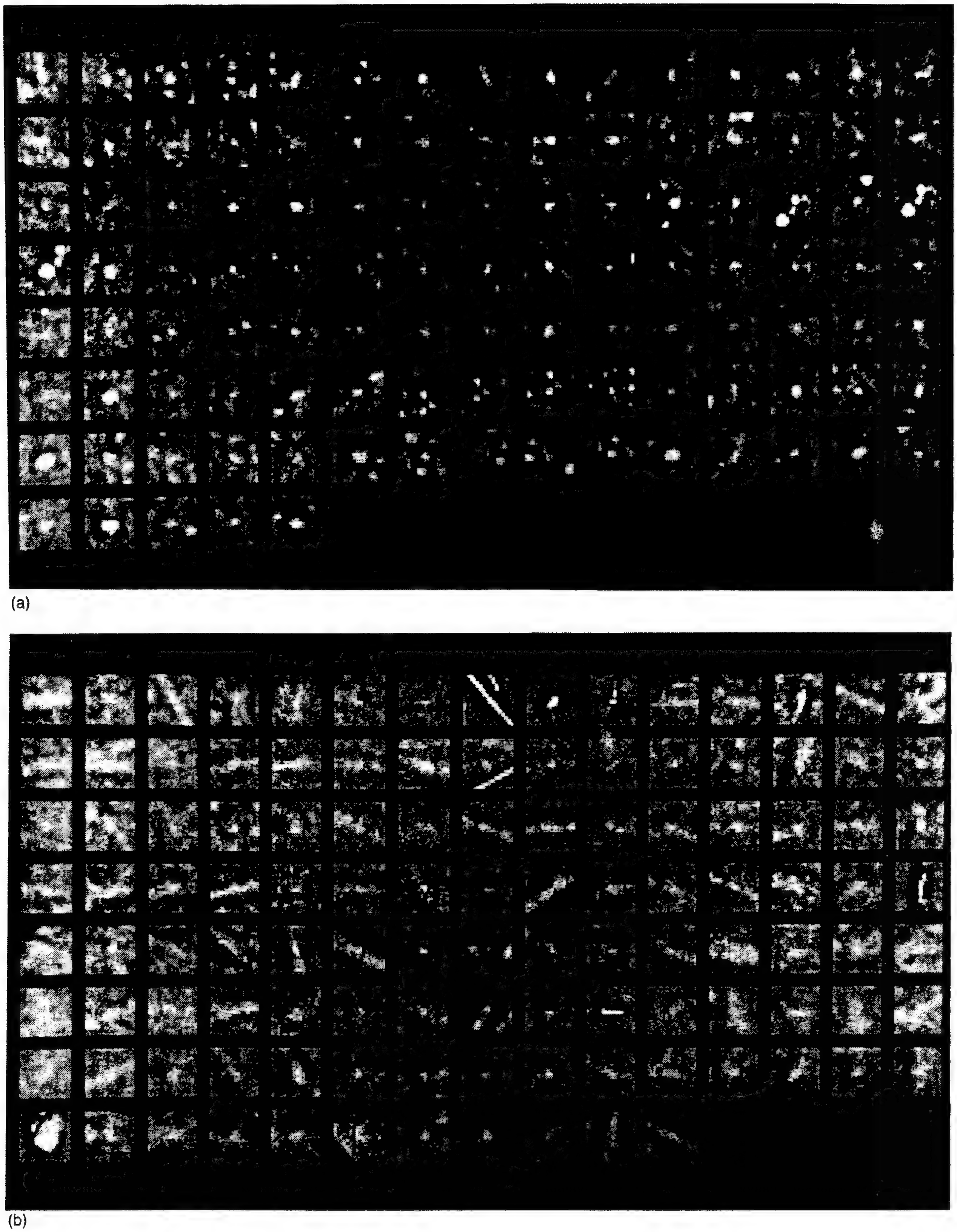
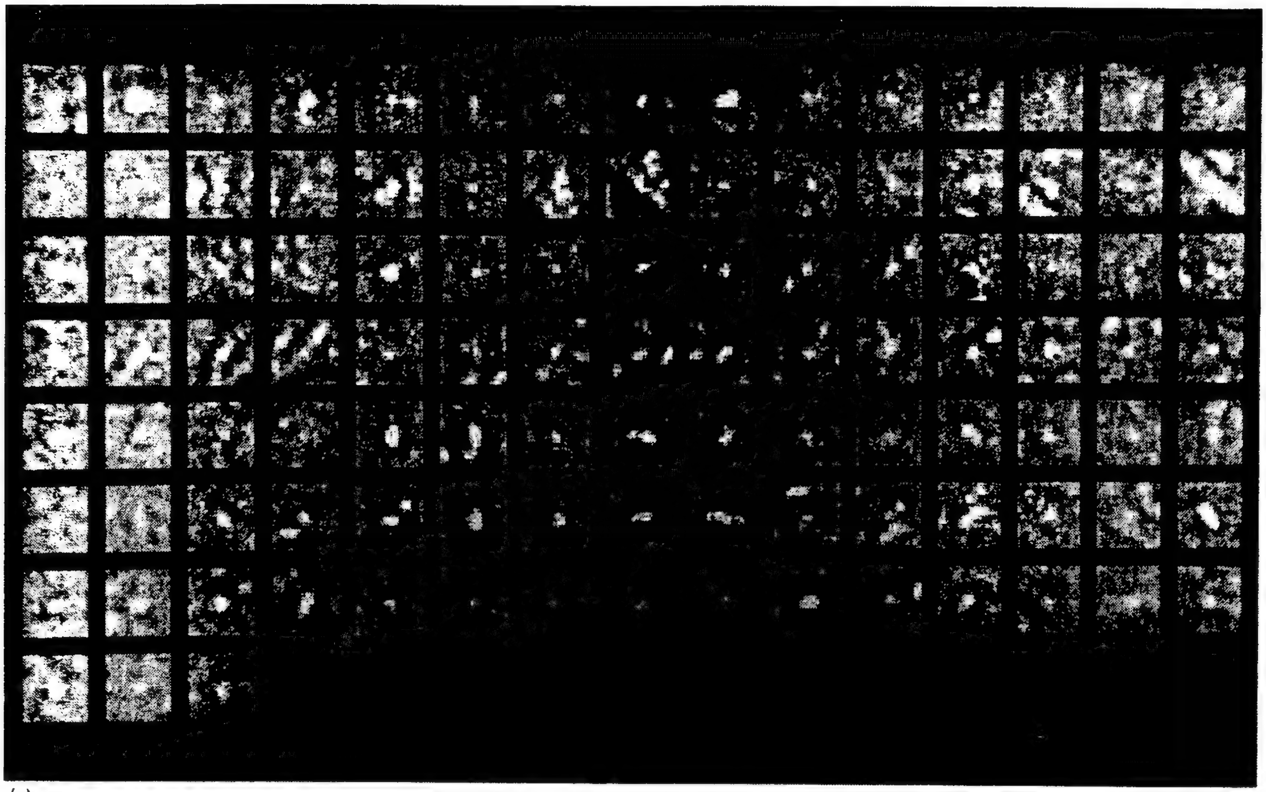


FIG. 2. The two groups of ROIs with true microcalcifications and false positives used for training of the CNNs in this study. Each of the ROI shown here contains 16×16 pixels (1.6 mm×1.6 mm). (a) ROIs in group 1 with true microcalcifications. (b) ROIs in group 1 with false positives. (c) ROIs in group 2 with true microcalcifications. (d) ROIs in group 2 with false positives.



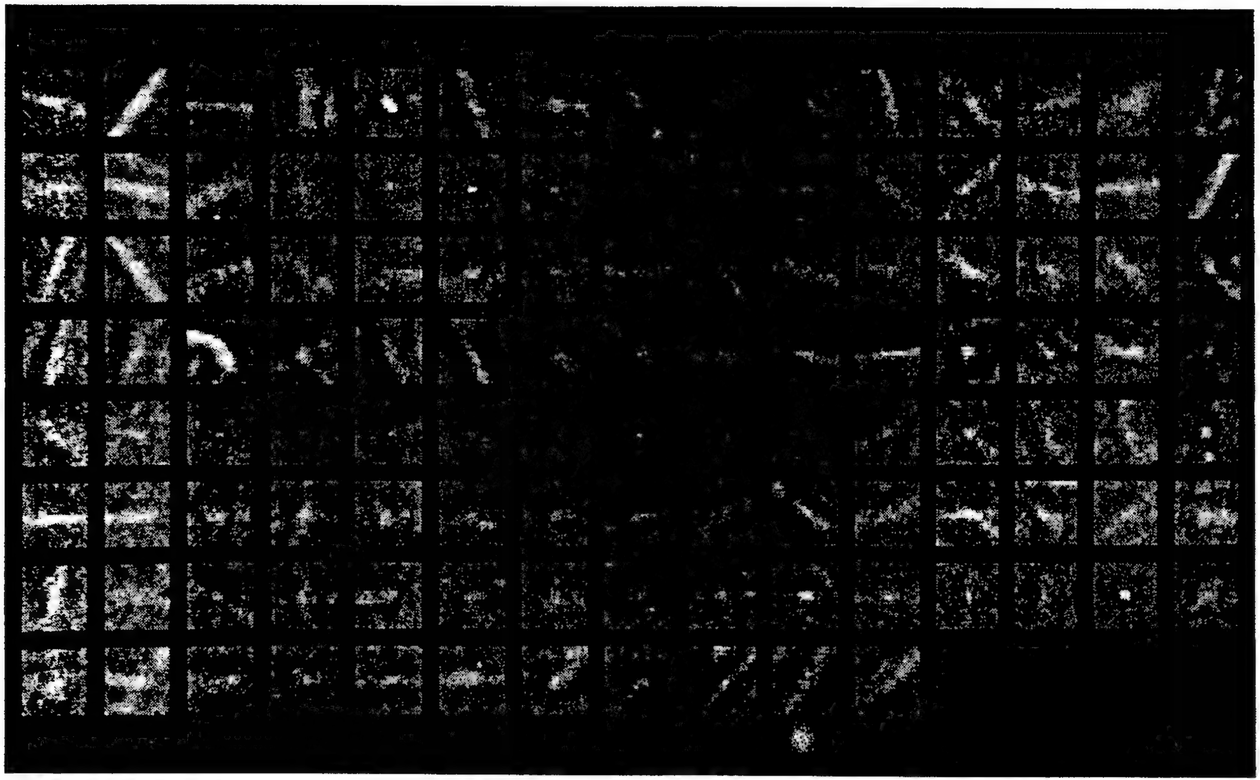


FIG. 2 (*Continued.*)

into the corresponding node in the upper layer. This value is further processed by the node through an activation function and produces an output signal that will, in turn, be forward propagated to the subsequent layer in a similar manner. The convolution kernel incorporates the neighborhood information in the input image pattern and transfers the information to the receiving layers, thus providing the pattern recognition capability of the CNN. The activation function between two layers is a sigmoidal function, and the signal at the Lth layer is obtained from the signal at the $(L-1)$th layer using the following relationship:

$$S_L((i,j);n) = \frac{1}{1+\exp\{-\Sigma_{\forall k\Rightarrow n}[w_L((i,j);k\Rightarrow n)*S_{L-1}((i,j);k)]\}}, \tag{1}$$

where $S_L((i,j);n)$ denotes the signal at node (i,j) in the nth group and Lth layer, $w_L((i,j);k\Rightarrow n)$ denotes the weight kernel connecting the kth group in the $(L-1)$th layer to the nth group in the Lth layer, * denotes the convolution operation, and the summation is over all groups k that are connected to group n. Note that the weight kernel for a given k and a given n is shift invariant, such that

$$w_L((i',j';i,j);k\Rightarrow n)=w_L((i'-i,j'-j);k\Rightarrow n), \tag{2}$$

where (i',j') denotes the node in the kth group and the $(L-1)$th layer. Because of the convolution operation, the useful matrix size of a node group in the Lth layer, $N_L\times N_L$, is reduced to $(N_{L-1}-K_{L-1}+1)\times(N_{L-1}-K_{L-1}+1)$, where $N_{L-1}\times N_{L-1}$ is the matrix size of a node group in the $(L-1)$th layer and $K_{L-1}\times K_{L-1}$ is the size of a weight kernel between the Lth layer and the $(L-1)$th layer.

In the output layer, there are n_{out} individual output nodes. Each output node is fully connected to all nodes in each group of the preceding hidden layer. The signal at the nth output node is given by Eq. (1), in which the weight matrix size is the same as the group size in the preceding layer and the output group size is 1×1.

F. Back-propagation training

The error back-propagation learning rule is used for supervised training of the CNN. The error function that is to be minimized by training is given by

$$\text{Error}=\frac{1}{2}\sum_{i=1}^{n_{\text{out}}}[S_{\text{in}}(i)-S_{Lo}(i)]^2, \tag{3}$$

where $S_{\text{in}}(i)$ is the input (or desired) value of a given training case at the ith node of the output layer, L_o, $S_{Lo}(i)$ is the network output signal of the case at that node, and n_{out} is the number of nodes in the output layer.

The conventional steepest descent delta rule for back-propagation training of a CNN can be written as

$$\begin{aligned} w_L((u,v);k\Rightarrow n)[t+1] \\ &= w_L((u,v);k\Rightarrow n)[t]+\eta\sum_{i,j}\delta_L((i,j);n) \\ &\quad \times S_{L-1}((i+u,j+v);k), \end{aligned} \tag{4}$$

where t is the number of iterations, η is the learning rate, and δ_L is the weight-update function given by

$$\delta_L((i,j);n)=S_L((i,j);n)[1-S_L((i,j);n)]Q_L((i,j);n), \tag{5}$$

where

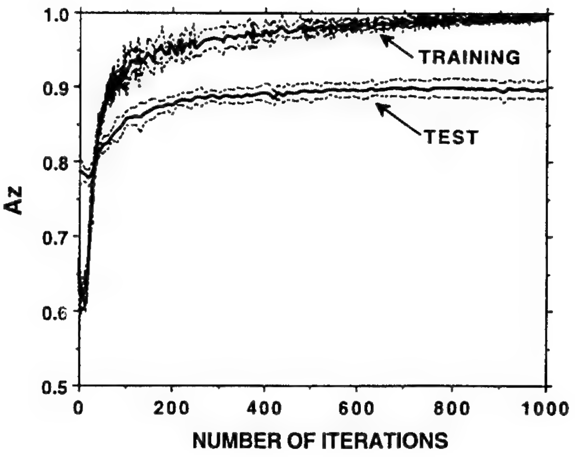


FIG. 3. Dependence of classification accuracy, A_z, on the number of iterations. The solid curves are the average A_z obtained from four repeated runs. The two dotted curves around each solid curve indicate the average±one SD of A_z estimated from the repeated runs. CNN configuration: 16×16 input nodes, first hidden layer: 12 node groups, second hidden layer: 12 node groups, each connected to 8 groups in the first hidden layer, two output nodes, weight kernels between input and first hidden layer: 5×5, weight kernels between first and second hidden layers: 3×3.

$$\begin{aligned} Q_L((i,j);n) &= \sum_{u,v;\forall k\Rightarrow n} w_{L+1}((u,v);k\Rightarrow n) \\ &\quad \times \delta_{L+1}((i-u,j-v);k\Rightarrow n). \end{aligned} \tag{6}$$

At the output layer, the weight is updated as

$$\begin{aligned} w_{Lo}((i,j);k\Rightarrow n)[t+1] &= w_{Lo}((i,j);k\Rightarrow n)[t] \\ &\quad +\eta\delta_{Lo}(k)S_{Lo-1}((i,j);n), \end{aligned} \tag{7}$$

where

$$\delta_{Lo}(k)=S_{Lo}(k)[1-S_{Lo}(k)][S_{\text{in}}(k)-S_{Lo}(k)]. \tag{8}$$

Training may be terminated at a selected level of total error, which is the sum of the error for an individual case [Eq. (3)] over all cases in the training set, a selected level of classification accuracy (A_z) as defined below, or a preset number of iterations. In this study, we used the total error as the termination criterion. The total error allowed at termination was chosen to be low enough so that the test A_z could reach a plateau, as demonstrated in Fig. 3.

In our application, all weights in the CNN were initialized to be between -0.5 to $+0.5$ using a random number generator with a different seed in each training run and normalized by the number of weights in the exponential factor of the sigmoidal activation function [Eq. (1)]. An $N\times N$-pixel region centered at a potential site that passes the size and contrast tests formed the input ROI to the CNN. For a given input SNR threshold, the program would identify a number of potential signals. A low SNR threshold corresponded to a lax criterion with a large number of false-positive (FP) signals. A high SNR threshold corresponded to a stringent criterion with a small number of FP signals and a loss in true-positive (TP) signals. For training the CNN, we arbitrarily divided the 38 mammograms in the average and subtle groups into two subgroups. When the ROIs obtained from one subgroup were used for training, the trained CNN would

TABLE I. Physical characteristics of microcalcifications extracted with an SNR threshold of 2.0 from the three groups of unfiltered mammograms.

Image group	No. of images	No. of μcalc.	Mean no. of μcalc/image	Size (pixels)		Contrast (pixel value)		SNR	
				Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Ratings 1,2	14	213	15	12.3	11.4	183.4	84.2	5.8	2.6
Rating 3	16	162	10	13.4	12.5	164.6	82.8	5.5	2.6
Ratings 4,5	22	270	12	9.0	9.4	143.9	87.0	4.6	2.2

be applied to the second subgroup for testing, and vice versa. We chose one of the SNR thresholds that yielded a moderate number of FPs and a sufficiently large number of TPs for segmenting the training ROIs. Because the number of FPs were still a few times more than the number of TPs, a subset of FPs with approximately the same number as the TPs were randomly chosen for the training set. It should be noted that the chosen SNR threshold level was not critical as long as the numbers of FP and TP were sufficiently large to provide the variety of ROI patterns for training the CNN. The ROIs obtained by using a high SNR threshold were generally a subset of those obtained by using a low SNR threshold. A chosen ROI input to the CNN was obtained from the SNR-enhanced image. The gray level values of the pixels in the ROI were thus independent of the SNR threshold at which it was chosen, and all ROIs had the same average background pixel value.

The shape of the microcalcifications in the breast parenchyma could be considered randomly oriented if we considered all possible locations of the microcalcifications in the breast and all mammographic views. To increase the variability of the training group, eight input ROIs to the CNN were generated from each ROI by rotating the ROI and its mirror image to 0°, 90°, 180°, and 270°. Each training cycle thus included training of the complete set of training ROIs with the eight orientations. The input order of the training ROIs was randomized with a different random number sequence in each run. A test ROI would be rotated also in the eight orientations, and the average output value of the eight rotated ROIs was taken to be the output value of that test ROI. During training, the desired output of an ROI with microcalcification was set to 1 and that of an ROI without microcalcification was set to 0.

We investigated the dependence of the classification accuracy of positive and negative ROIs on the CNN configurations. Because of the computational requirements in training the CNNs, we did not exhaustively study every possible combination of parameters. The range of parameters that we studied and the corresponding results are tabulated in Table III. CNNs with one and two hidden layers were examined. The number of node groups in the hidden layers was varied from 4 to 12. In most of the two-hidden-layer CNNs, the number of groups was kept the same for both layers. Combinations of 12 groups in the first hidden layer and 4, 8, or 12 groups in the second hidden layer were also studied. All node groups in the two hidden layers are fully connected in these configurations. Additionally, a 12 group–12 group combination in which every 3 of the 12 groups in the second hidden layer were connected to the same 8 selected groups in the first hidden layer was examined.[34] For comparison, a CNN with 8 groups in the first hidden layer and 12 groups in the second hidden layer with full connections was included. We also evaluated the classification accuracy for two combinations of weight kernel sizes; one had a kernel size of 5×5 in the first hidden layer and 3×3 in the second hidden layer and the other had a kernel size of 7×7 in the first hidden layer and 5×5 in the second hidden layer. We did not investigate larger kernel sizes because the sizes of the microcalcifications of interest were generally much smaller than 7×7 pixels and because computation time increased rapidly with kernel size. The input ROI size was adjusted so that the size of the node groups in the last hidden layer was 10×10 for both combinations of kernel sizes. The output nodes were always fully connected to every node group in the last hidden layer with a 10×10 kernel, as shown in Fig. 1.

TABLE II. Number of ROIs with microcalcifications and false positives for training of the CNN.

	Number of ROIs			
	Group 1 (G1)	with rotation	Group 2 (G2)	with rotation
Microcalcifications	110	880	108	864
False positives	116	928	116	928

The classification accuracy of the CNN during training was monitored by receiver operating characteristic (ROC) analysis[38] of the output values from the CNN. After each iteration, or epoch, with the training set was completed, the classification performance with the current weights for all training cases would be determined by inputting the training cases into the CNN as a consistency verification procedure. The distributions of the output values for the positive ROIs and the negative ROIs would be input into the LABROC1 program,[39] which assumes binormal distributions of the decision variable for the normal and abnormal cases and fits an ROC curve based on maximum likelihood estimation. The ROC curve represents the relationship between the true-positive fraction (TPF) and the false-positive fraction (FPF) as the decision threshold varies. The LABROC1 program provides the area under the fitted ROC curve, A_z, and an estimate of the SD of A_z. A_z is used as an index of classification accuracy. The dependence of A_z on the number of iterations was monitored during training. For every ten training iterations, the trained CNN was applied to the other independent set of ROIs to test its generalization capability. The dependence of the test A_z on the number of iterations was also examined.

TABLE III. Dependence of test results, in terms of the area under the ROC curve (A_z), on the configuration of CNN for classification of microcalcifications.

Input ROI size (pixels)		16×16			20×20		
Kernel size (first hidden layer)		5×5			7×7		
Kernel size (second hidden layer)		3×3			5×5		
No. of groups in hidden layer		No. of output nodes			No. of output nodes		
First	Second	1	2	2	1	2	2
		A_z Train: G1 Test: G2	A_z Train: G1 Test: G2	A_z Train: G2 Test: G1	A_z Train: G1 Test: G2	A_z Train: G1 Test: G2	A_z Train: G2 Test: G1
4	4	0.86	0.86	0.86	0.90	0.87	...
6	6	0.88	0.88	0.86	0.89	0.90	0.89
8	8	0.88	0.88	0.88	0.89	0.90	0.89
10	10	0.89	0.89	0.88	0.91	0.90	0.89
12	12	...[a]	0.91	0.89	...	0.91	0.89
12	4	...	0.88	0.86	...	0.90	0.90
12	8	0.89	0.90	0.89	0.90	0.90	0.90
8	12	...	0.89	0.89	...	0.88	0.90
12(8)[b]	12	0.90	0.90	0.90	0.91	0.90	0.89
One hidden layer							
4		0.87	0.85	...	0.83	0.85	...
8		0.85	0.87	...	0.86	0.85	...
12		0.87	0.86	...	0.86	0.86	...

[a] The CNN configuration was not tested if there is no entry.

[b] Eight node groups in the first hidden layer are selectively connected to the 12 node groups in the second hidden layer. The A_z values for this CNN are the averages of four runs shown in Table IV.

G. Analysis of detection accuracy

After passing the size and contrast criteria, being screened by the trained CNN, and passing the regional clustering criterion, the detected individual microcalcifications and clusters would be compared with the "truth" file of the input image. The number of TP and FP microcalcifications and the number of TP and FP clusters were scored. A detected signal was scored as a TP microcalcification if it was within 0.5 mm from a true microcalcification in the "truth" file. A detected cluster was scored as a TP if its centroid coordinate was within a cluster radius (5 mm) from the centroid of a true cluster and at least two of its member microcalcifications were scored as TP. Once a true microcalcification or cluster was matched to a detected microcalcification or cluster, it would be eliminated from further matching. Any detected microcalcifications or clusters that did not match to a true microcalcification or cluster were scored as FPs. The tradeoff between the TP and FP detection rates by the computer program was evaluated by the free-response receiver operating characteristic (FROC) analysis[40] by varying the input SNR threshold. A low SNR threshold corresponded to a lax criterion with a large number of FP clusters. A high SNR threshold corresponded to a stringent criterion with a small number of FP clusters and a loss in TP clusters. The detection accuracy of the computer program with and without the CNN classifier could then be assessed by comparison of the FROC curves.

III. RESULTS

Using the signal extraction program described in Sec. II, the size, contrast, and SNR of the true microcalcifications as indicated in the "truth" file for each of the three groups of mammograms were determined at several SNR thresholds. We examined the extracted signals in the thresholded images and compared visually the extracted signals with those in the original images. When the SNR threshold was too low, the signals merged with one another or with noise in the background. The extracted signals did not represent the true signal size or shape. When the SNR threshold was too high, many subtle microcalcifications were not extracted. The extracted signals appeared to be smaller than those in the original images because only a few pixels in a microcalcification were higher than the threshold. It was determined subjectively that an SNR threshold of 2.0 was a compromise with which the extracted signals were similar in size and shape to those in the original images. At this SNR threshold, an average of about 85% of the microcalcifications were extracted. The other 15% of the microcalcifications could not be extracted at this threshold because their pixel values were lower than the local gray level threshold.

Table I shows the mean and SD of the contrast, size, SNR of the microcalcifications extracted at an SNR threshold of 2.0 for each of the three groups. Note that the "size" of an extracted microcalcification depends on the SNR threshold used because it may merge with an adjacent noise or signal pixels, as discussed previously. The contrast is relatively independent of the SNR threshold, since it depends only on the maximum pixel value in the signal region. We have plotted the histograms of the contrast, size, and SNR of the extracted microcalcifications and found a large overlap in the physical characteristics of the microcalcifications in the three groups of mammograms. As can be seen in Table I, the visual ranking generally correlates with the mean contrast and mean SNR of the microcalcifications. However, the mean number of microcalcifications in the subtle group is larger than that of the average group. These observations indicate that the visibility of a microcalcification cluster is more strongly affected by the contrast and SNR than by the number of microcalcifications in the cluster. This is consistent with the experience of radiologists in visual detection of microcalcifications. The data in Table I should provide more objective information than the visibility ratings in the description of the degree of subtlety for each group of microcalcifications. The quantitative characterization can facilitate comparison of the performance of CAD algorithms in different datasets if a similar signal extraction method and criteria are used in calculation of the data.

Table II shows the number of ROIs with true microcalcifications and false signals used for training of the CNN. Each group of ROIs was detected with the automated algorithm at an SNR threshold of 3.4 from 19 SNR-enhanced images. At this threshold, the average TP rate was 94% at an average FP rate of 7.5 clusters per image. This point was outside the range of the FP rates plotted in Fig. 7. There were no overlapping cases in the two groups. The extracted ROIs of 16×16 pixels are displayed in Figs. 2(a)–2(d). It can be seen that a large number of the FPs extracted by the CAD program was caused by high-frequency structures such as fibrous strands, film artifacts, and noise. Only one-fifth of the FP ROIs were included in the training groups in order to match approximately the number of ROIs with true microcalcifications. With the rotation method, over 800 positive and over 900 negative ROIs were generated in each training set. When one group was used for training, the displayed ROIs in the other group, together with the other four-fifths of the FP ROIs from the same set of images, were used as the test set. The signal of interest was centered at the ROI. The average background gray level was the same for all ROIs after the SNR-enhancement filtering.

The dependence of the classification accuracy on CNN configuration and training set is shown in Table III. The SDs of the A_z as determined by the LABROC1 program ranged from 0.01 to 0.02. The classification accuracy during training generally reached an A_z of 0.99 or greater under all conditions studied. The test results exhibited some variations, as can be seen from Table III. The CNNs with one hidden layer are inferior to the CNNs with two hidden layers. The performance of the CNNs with two hidden layers does not depend strongly on the configuration when the total number of weights in the CNN is large. There is a slight trend, with some minor variations, that the A_z increases as the number of node groups increases. This trend is more systematic for the CNNs with small weight kernels. There is also a trend that, for the same CNN configuration, the test A_z is larger when G1 is used for training than when G2 is used. The difference in the test A_z values between the two training/test group combinations, averaged over all two-hidden-layer, two-output-node CNNs studied, is only about 0.01. This difference, however, is statistically significant at a two-tailed p level of 0.005.

TABLE IV. Reproducibility of test results for two CNNs. The A_z shown is the maximum value reached for a given run.

Input ROI size (pixels)	16×16		20×20	
Kernel size				
First hidden layer:	5×5		7×7	
Second hidden layer:	3×3		5×5	
No. of groups				
First hidden layer:	12(8)		8	
Second hidden layer:	12		8	
Repeated run	Train: G1 Test: G2	Train: G2 Test: G1	Train: G1 Test: G2	Train: G2 Test: G1
	A_z		A_z	
1	0.91	0.91	0.91	0.90
2	0.90	0.88	0.90	0.88
3	0.89	0.89	0.90	0.89
4	0.90	0.91	0.90	0.88
Mean	0.90	0.90	0.90	0.89
std. dev.	0.01	0.01	0.01	0.01

We also compared the difference in performance between CNN with one- and two-output nodes. The two output values from the two-output CNNs were found to be complementary to each other, i.e., for a given case, if the output of one node was x, the output from the other node was very close to $(1-x)$. This outcome is expected because the desired output values for the true and false signals were set to be 1 and 0, respectively, as described above. Therefore, the output from one node was sufficient for the classification task, and the ROC curve could be constructed from either of the output nodes. The difference in the A_z values between the one- and two-output configurations, averaged over the two-hidden-layer CNNs, is 0.001. The difference is not statistically significant ($p=0.68$).

To study the variability in the classification accuracy due to the initialization condition of the weights and training, the training and testing of two selected CNN configurations were repeated four times for each of the two training/test group combinations. The CNN configurations and the results are listed in Table IV. The SD of A_z is estimated from the repeated runs to be 0.01 in each case, and the maximum difference in A_z for the four runs is 0.03. For a given CNN configuration, the mean A_z values for the two training/test combinations agree within 0.01. Figure 3 shows the dependence of A_z for training and testing, averaged over four runs, on the number of iterations for one of the CNNs. The SDs of A_z estimated from the repeated runs are also plotted for the training and the test curves. Both the mean A_z values and SDs stabilize after some large fluctuations in the initial iterations. The shapes of the A_z curves are typical of the conditions included in this study, although the rate of convergence varies with CNN configurations. The curves increase rapidly initially then plateau off and gradually approach its maximum level. The convergence of the CNN training can also be observed from the dependence of the total error on the number of iterations, as shown in Fig. 4. The magnitude of the error depends on the number of output nodes and the number of input cases. However, the trend of the curve is typical among the CNNs studied. It shows a steep descent initially

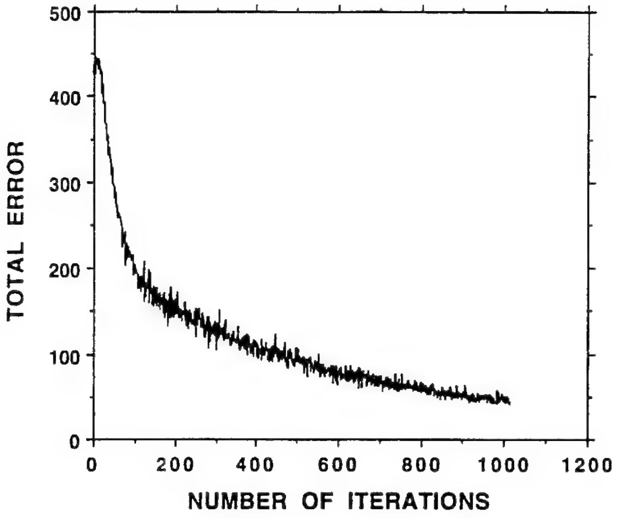


FIG. 4. Dependence of total error of the CNN output on the number of iterations. The CNN configuration is the same as that in Fig. 3. Training group G2 was used.

then gradually levels off at large number of iterations.

The training of each CNN with each group of training cases produces a set of weights at each iteration. Many of the CNN configurations reach approximately the same level of performance (Table III) and may be used as a classifier in the microcalcification detection program. We selected one of the trained CNNs shown in Table IV (2 hidden layers, each with 12 node groups, every 3 node groups in the second hidden layer selectively connected to 8 node groups in the first hidden layer, weight kernels sizes of 5×5 and 3×3) to demonstrate the effect of the CNN classifier on detection accuracy. A weight set trained with the G1 group was used to test the classification accuracy for the G2 group and another set trained with the G2 group was used to test the classification accuracy for the G1 group. The weights were obtained from one of the iterations when the plateau of A_z was reached. The ROC curves for classification of the test groups of ROIs using the trained CNNs are shown in Fig. 5. The A_z values of the curves are 0.91 and 0.90, which correspond to the best performance obtained with the CNNs tabulated in Table III.

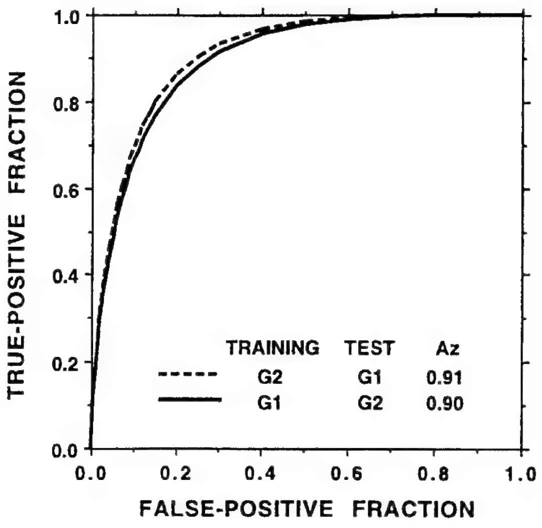


FIG. 5. The ROC curves obtained with the test ROI groups. The CNN configuration is the same as that in Fig. 3.

We incorporated the trained CNN into our microcalcification detection program as described previously, and the overall improvement in the detection accuracy was evaluated. For any SNR threshold, each extracted signal that passed the size and contrast criteria was input into the CNN. The set of weights obtained from training with G1 was used for the 19 mammograms from which the G2 ROIs were extracted, and vice versa. For the group of obvious mammograms, either set of weights could be used because the obvious cases were not used for training or testing. The performance of the trained CNNs on the obvious cases was thus an additional independent test for the classifiers. In this application, a constant decision threshold was set for the CNN output value of any input ROI to determine if the ROI was normal or abnormal. To select the appropriate decision threshold for the output value from the CNN, the dependence of the FROC curve on the decision threshold was evaluated. This corresponded to varying the operating point along the ROC curve (Fig. 5) of the classifier. The FROC curves for the three sets of mammograms were plotted in Figs. 6(a)–6(c). The data points along each FROC curve were obtained by varying the SNR thresholds from 3.0 to 5.2. Some of the data points were not plotted if they were outside the range of the graph. A curve without the CNN (decision threshold=0), and two curves with CNN at decision thresholds of 0.5 and 0.8, respectively, were plotted. For a given TP rate, the number of FP clusters decreased as the CNN threshold increased from 0.1 to 0.8. When the CNN threshold was further increased to 0.9, we observed a decrease in the TP rate for a given FP rate for subtle cases, indicating that many of the ROIs with subtle microcalcifications were misclassified with the high CNN threshold. At a CNN threshold of 0.8, the TP rate was 100% at an FP rate of less than 0.1 cluster per image for the obvious cases. For the cases that were ranked average subtle by radiologists, the TP rate was about 93% at an FP rate of one cluster per image. For the subtle cases, the TP rate was 87% at an FP rate of 1.5 clusters per image.

The degree of subtlety of the clustered microcalcifications in the cases ranked from 3 to 5 is similar to that of the cases used in our previous observer performance study.[11] The likelihood that these microcalcifications may be missed is not negligible, and thus it is of particular interest for CAD applications. The average improvement in the detection accuracy for these cases is estimated by comparison of the FROC curves without and with the CNN classifier for all cases ranked 3–5. The FROC curves are shown in Fig. 7. The TP rate improves from about 87% at an FP rate of 4 clusters per image without the CNN classifier to 90% at an FP rate of about 1.5 clusters per image, with the CNN classifier at a decision threshold of 0.8.

IV. DISCUSSION

The computational cost for training a CNN is high. The computational cost per iteration increases as the numbers of nodes and weights increase. However, it was observed that the rate of convergence increased as the number of nodes

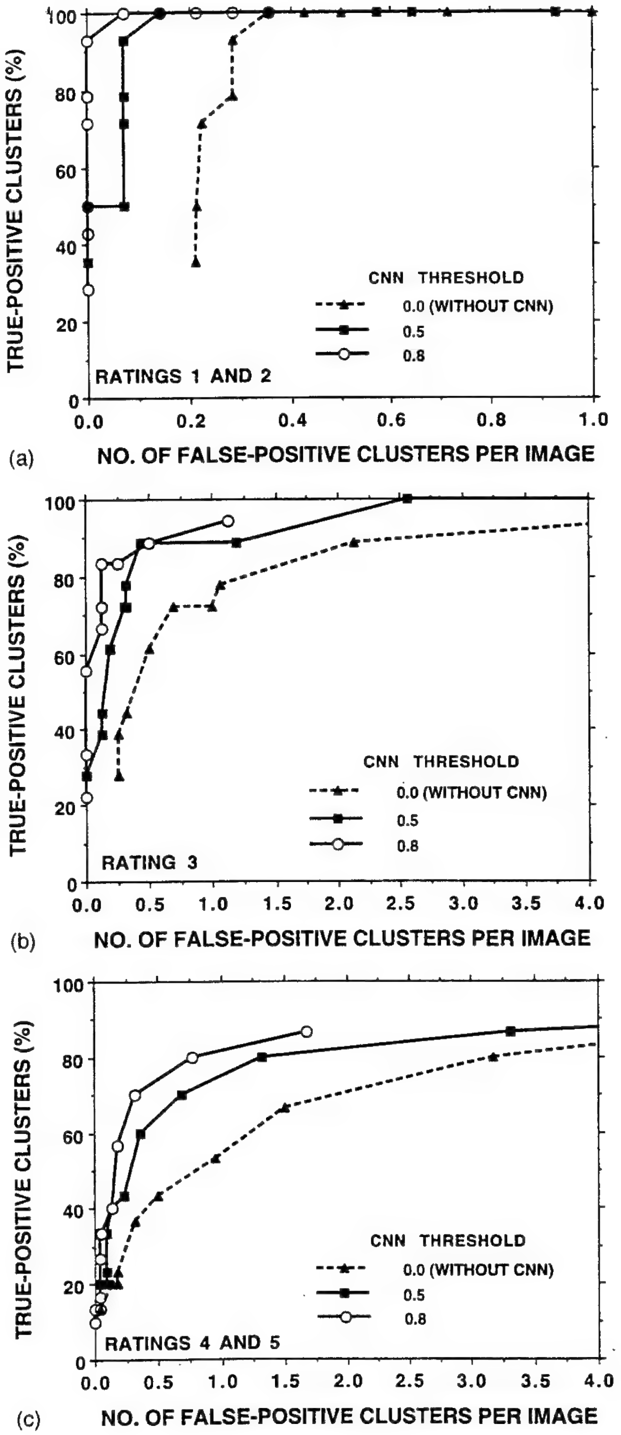


FIG. 6. Comparison of FROC curves for detection of clustered microcalcifications without and with the CNN classifier. The curve without CNN is equivalent to that with the decision threshold of the CNN set to 0. The FROC curves with the decision threshold of the CNN set to 0.5 and 0.8 are plotted for comparison. (a) Mammograms with obvious microcalcifications. (b) Mammograms with average subtle microcalcifications. (c) Mammograms with subtle microcalcifications. The CNN configuration is the same as that in Fig. 3.

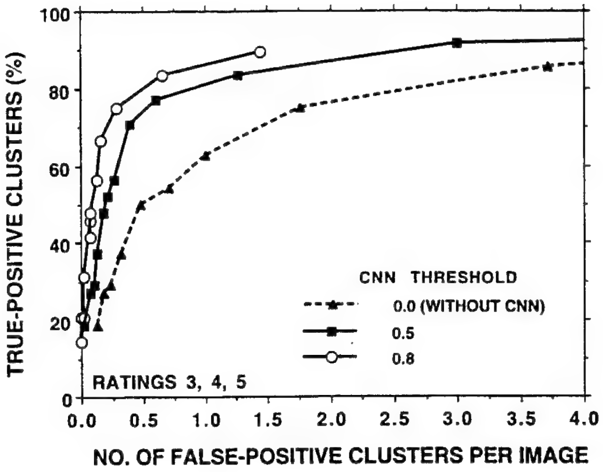


FIG. 7. Comparison of FROC curves for detection of clustered microcalcifications without and with the CNN classifier. The overall detection accuracy for the average and subtle groups of microcalcifications are compared. The CNN configuration is the same as that in Fig. 3.

increased. For example, for CNNs of the same configuration, except for a difference in the number of output nodes, the two-output-node CNN reached the maximum A_z with a smaller number of iterations than the corresponding one-output-node CNN. This trend is more obvious for the CNNs with fewer node groups in the hidden layers. Similarly, the convergence rate increases until the number of node groups in the hidden layers increases to about 10 for CNNs with two-output nodes.

The convergence rate saturates sooner for the CNNs with a larger kernel size. Therefore, the overall training cost of CNNs with complicated configurations may not be higher than those with simpler configurations. We could not perform an exact comparison of the computation time for different CNN configurations because we had to make use of all available workstations that had different CPU speeds and different memory capacities to train the CNNs. It may be noted that the computational cost with a complicated CNN configuration is higher than that of a simple one when it is incorporated in the microcalcification detection program for classification of test cases.

We have attempted to apply the FROCFIT curve fitting program[41] to the FROC curves in this study, but failed to obtain well-fitted curves. This may be caused by the fact that the FROCFIT was developed on the basis of several assumptions, which may not be satisfied for our detection task.[35] We therefore could not arrive at a single value such as the A_1[41] as the performance index for comparison of the different conditions. The generalization capability of the CNN can be observed from the effectiveness of the trained CNN in reducing FPs in the additional independent test group of mammograms of ratings 1 and 2 [Fig. 6(a)]. At a CNN threshold of 0.8, the FP clusters were reduced to zero for almost all TP rates below 100%. Because there is no established method to test the statistical significance of the difference in two FROC curves, we performed t tests on the image-specific paired FP values between the without-CNN and with-CNN (threshold $=0.8$) results at corresponding TP rates, in an effort to estimate the significance of their differences. The p values ranged from 0.04 to 0.08. Although the improvement in the FP rates was very consistent over the entire range of TP rates, as shown in Fig. 6(a), the level of significance for individual TP rates was not high, probably because the FP rates without CNN were already very low.

The performance of the trained CNN can also be observed from the effective reduction of FPs at different SNR thresholds in the test group of mammograms of ratings 3–5. As shown in Fig. 7, for a given TP rate, the CNN reduced the FP clusters by more than 70% with a CNN threshold of 0.8. We again performed t tests on the image-specific paired FP values at corresponding TP rates. For TP rates between about 20%–75% the p values of the differences between the FP rates without CNN and with CNN (threshold=0.8) ranged from 0.06 to 0.0002. We also performed t tests on the paired TP rates at corresponding FP rates. For FP rates between about 0.1 to about 0.7 clusters per image, all p values of the differences between the TP rates without CNN and with CNN (threshold=0.8) were less than 0.001.

The FROC curves presented here were obtained by varying the SNR threshold in the local gray level thresholding process. The CNN classifier was implemented so that a constant decision threshold for its output value was used to classify ROIs with and without microcalcifications obtained at any SNR threshold. Alternatively, we can select a relatively low SNR threshold that produces a large number of FPs and vary the decision threshold for the output of the CNN classifier, thereby generating pairs of TP and corresponding FP values along an FROC curve. We have studied this approach by using SNR thresholds from 3.0 to 5.2, from each of which an FROC curve was generated by varying the CNN threshold from 0.1 to 0.9. It was observed that the FROC curves obtained with this alternative method were lower than the FROC curve with CNN (threshold=0.8) plotted in Fig. 7. On each of these alternative FROC curves, the data point at a CNN threshold of 0.8 coincided with the data point on the FROC curve with CNN (threshold=0.8) shown in Fig. 7, because they are the data points with the same SNR and CNN thresholds. Other data points on the alternative FROC curves are either comparable to or lower than the FROC curve in Fig. 7, with a few exceptions in the range of very low TP and FP rates.

The goal of this study is to evaluate the feasibility of training a CNN to distinguish FP signals from true microcalcifications obtained from our automated detection program. Although a small dataset was used and sample biases may exist, the effectiveness of the method as one of the steps in the classification process was demonstrated by the relative improvement in the detection accuracy. In the field of CAD, it is known that different detection algorithms or even different human observers may generate FPs of different characteristics. Before the CNN classifier is to be incorporated into a CAD program for clinical implementation, it is important to train the classifier using true and false microcalcifications obtained from the specific application. The training dataset should also be large enough to ensure that the patient population is adequately represented and that the performance of the classifier can be generalized.

V. CONCLUSION

We have developed a computer program for automated detection of clustered microcalcifications on mammograms for CAD applications. In this study, we investigated the effectiveness of a new signal classifier based on artificial neural network methodology for improvement of the detection accuracy of the CAD program. The CNN classifier was trained to recognize individual microcalcifications and incorporated as one of the signal classification steps. It was found that the CNN classifier can achieve a classification accuracy, expressed in terms of the A_z index with ROC analysis, of 0.9. It reduced the average FP rates by more than 70% at all TP rates on mammograms with subtle to obvious microcalcifications. Although the number of cases used in this study is limited, the improvement is consistent and statistically significant. This study demonstrates that a CNN can be trained to recognize mammographic microcalcifications and is effective in reducing FP detections in CAD applications.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant No. CA 48129 and U.S. Army Grant No. DAMD 17-93-J-3007 (through subgrant No. GU RX 4300-803UM from Georgetown University). The content of this publication does not necessarily reflect the position of Georgetown University or the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC1 programs, to Dev Chakraborty, Ph.D., for the FROCFIT program, and to Diane Williams for secretarial assistance.

[a] Radiology Department, Imaging Physics Division, Georgetown University, Washington, DC 20007.

[b] Department of Radiation Oncology, University of Michigan.

[1] *National Center for Health Statistics. Vital Statistics of the United States, 1987. Vol. 2. Mortality. Part A*, DHHS Publication No. (PHS) 90-1101 (Government Printing Office, Washington, DC, 1990).

[2] J. R. Harris, M. E. Lippman, U. Veronesi, and W. Willett, "Breast cancer," N. Engl. J. Med. **327**, 319–328 (1992).

[3] M. Moskowitz, "Breast cancer: Age-specific growth rates and screening strategies," Radiology **161**, 37–41 (1986).

[4] M. Moskowitz, "Benefit and risk," in: *Breast Cancer Detection: Mammography and Other Methods in Breast Imaging*, edited by L. W. Bassett and R. H. Gold, 2nd ed. (Grune and Stratton, New York, 1987).

[5] H. Seidman, S. K. Gelb, E. Silverberg, N. LaVerda, and J. A. Lubera, "Survival experience in the breast cancer detection demonstration project," Cancer J. Clin. **37**, 258–290 (1987).

[6] J. E. Martin, M. Moskowitz, and J. R. Milbrath, "Breast cancer missed by mammography," Am. J. Roentgenol. **132**, 737–739 (1979).

[7] M. G. Wallis, M. T. Walsh, and J. R. Lee, "A review of false negative mammography in a symptomatic population," Clin. Radiol. **44**, 13–15 (1991).

[8] R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," Radiology **184**, 613–617 (1992).

[9] J. A. Harvey, L. L. Fajardo, and C. A. Innis, "Previous mammograms in patients with impalpable breast carcinomas: Retrospective vs blinded interpretation," Am. J. Roentgenol. **161**, 1167–1172 (1993).

[10] E. L. Thurfjell, K. A. Lernevall, and A. A. S. Taube, "Benefit of independent double reading in a population-based mammography screening program," Radiology **191**, 241–244 (1994).

[11] H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).

[12] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for speculated lesions," Radiology **191**, 331–337 (1994).

masses. Of the malignant masses, 45 had spiculated margins. Of the benign masses, 6 were spiculated. The visibility of the masses ranged from subtle to obvious. The average size (length of the long axis) of the masses, as estimated by the radiologists, was 12.2 mm., and the standard deviation of the mass size was 4.5 mm. The mammograms were randomly divided into training and test groups, each of which contained 84 mammograms.

The mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel size of $100\mu m \times 100\mu m$ and 4096 gray levels. The light transmitted through the film was amplified logarithmically before analog-to-digital conversion. The digitizer had an optical density (OD) range of 0–3.5. It was calibrated so that the OD on film was linearly proportional to the output pixel value in the range of about 0.1 OD to 2.8 OD with a slope of 0.001 OD/pixel value. The slope of the calibration curve outside this range decreased gradually.

Four different ROIs, each with 256×256 pixels, were selected from each mammogram by a radiologist experienced in mammography. One of the selected ROIs contained the true mass which was identified by an experienced radiologist and verified by biopsy reports. The remaining three ROIs contained breast parenchyma that was presumed to be normal, with the first region containing dense tissue which could mimic a mass lesion, the second region containing mixed dense/fatty tissue, and the third region containing fatty tissue. An example of each of these ROIs is shown in Fig. 2.

B. Background Correction

The masses superimpose on structured background tissue in the ROIs. In most cases, this background tissue is not uniform over the ROI. For example, one side of the ROI may contain denser tissue than the other side, or, when the mass is close to the outer edge of the breast, one corner of the ROI may contain a non-breast region. This may reduce the detectability of the mass by a neural network. To reduce this non-uniformity, we developed a background correction method that estimated the background level based on the image intensity in a band of pixels surrounding the ROI.

We estimated a background image $B(i,j)$ from the original ROI as follows. For a given point (i_0, j_0) in the original image, we computed four averages, L, R, U, and D, inside the boxes A_L, A_R, A_U, and A_D shown in Fig. 3. The centers of the boxes were either at the same row or at the same column as (i_0, j_0). The box size was 16×32 if it could be placed entirely inside the ROI. Near the corners of the image, the box size was gradually decreased in order to avoid edge effects. For example, if the center of the box was exactly at a corner, then the box size was 16×16. The background pixel $B(i_0, j_0)$ was interpolated from the averages L, R, U, and D as

$$B(i_0,j_0) = \left[\frac{L}{d_l} + \frac{R}{d_r} + \frac{U}{d_u} + \frac{D}{d_d}\right] \Big/ \left[\frac{1}{d_l} + \frac{1}{d_r} + \frac{1}{d_u} + \frac{1}{d_d}\right], \tag{11}$$

where d_l, d_r, d_u and d_d are the distances between (i_0, j_0) and each side of the image. The background image was then subtracted from the original image, thus reducing the background to near 0.

As mentioned in the previous subsection, the average size of the masses was 12.2 mm, which corresponded to 122 pixels after digitization. Since the masses were placed in the center of the ROIs in the extraction process, very few of the ROIs contained mass tissue in the 16-pixel wide band that was used for background estimation. Out of 168 masses in our database, only four had a long axis longer than 220 pixels, and the long axis in these four cases was not exactly horizontal or vertical. We therefore believe that our estimation essentially excluded information about the mass itself, and included only information about the background. Fig. 4 shows an example of an ROI before and after background correction.

C. Classification with Subsampled Images

The simplest method of classifying mass and nonmass ROIs using a CNN would be to input the background-corrected images directly to the input layer of the CNN. However, the computational cost of inputting 256×256 ROIs into a CNN was prohibitive. We thus had to reduce the image size by averaging adjacent pixels and subsampling. We investigated the effect of reducing the image size to 16×16 and 32×32. Averaging was performed on non-overlapping boxes of size 16×16 to obtain 16×16 subsampled images, and of size 8×8 to obtain 32×32 subsampled images.

We investigated the use of a three-layer CNN with a single input image, and a single output node, as shown in Fig. 5. The number of hidden layer groups $N(2)$, and the weight kernel size between the input layer and the hidden layer $S_w(1)$, were variable. For this special case, CNN forward propagation equations simplified considerably. Let $H(i,j) \equiv H_{1,1}(i,j)$ denote the subsampled input image, $W_g(i,j) \equiv w_{1,g,1}(i,j)$ denote the weight kernel between the input layer and the g^{th} group in the hidden layer, and let $\mathcal{W}_{g'}(i,j) \equiv w_{2,1,g'}(i,j)$ denote the weight kernel between the g'^{th} group in the hidden layer and the output O. Then, the forward propagation equations (1)–(3) simplified to

$$I_{2,g} = H ** W_g, \qquad g = 1, \ldots, N(2), \tag{12}$$

$$H_{2,g}(i,j) = \frac{1}{1 + exp(-I_{2,g}(i,j))}, \quad g = 1, \ldots, N(2), \tag{13}$$

and

$$O = \frac{1}{1 + exp\left(-\sum_{g'=1}^{N(2)} H_{2,g'} ** \mathcal{W}_{g'}\right)}. \tag{14}$$

As mentioned in Section II.D., eight rotated and mirrored images were applied to the CNN consecutively for each averaged-subsampled ROI. The CNN output score for the ROI was obtained as the average of the CNN outputs for these eight images. The CNN output error for training image p was calculated as the square of the difference of the CNN output score and the desired CNN output. Backpropagation with the delta-bar-delta rule was implemented using equations (5)–(8).

After training, the averaged-subsampled images belonging to the test group were applied to the CNN with the trained weights, and the CNN test output scores were obtained using forward propagation. The CNN output scores were used as the decision variable in Receiver Operating Characteristics (ROC) analysis [24] to evaluate the classification performance. ROC analysis evaluates the relationship between the true-positive fraction (TPF) and the false-positive fraction (FPF) as the decision threshold varies. We estimated the ROC curve using the LABROC1 program [25] which assumes binormal distributions of the decision variable for the normal and abnormal cases and fits the ROC curve based on maximum likelihood estimation. The area under the ROC curve, A_z, was used as an index of classification accuracy. The classification results obtained with the CNN described in this subsection are presented in Section IV.A.

D. Classification with GLDS Texture-Images

D.1 GLDS Features

GLDS features, extracted from the GLDS vector of an image, roughly measure the coarseness of the texture elements in an image. The GLDS vector is the histogram of the absolute value of the difference of pixel pairs separated by a distance d_1 in the horizontal direction and d_2 in the vertical direction. The vector $d = (d_1, d_2)$ is called the displacement vector. As discussed below, the distribution of the elements of the GLDS vector $p_d(k)$ indicate size of the texture element in the image relative to the displacement vector d. GLDS features are extracted by computing some measure of the distribution of the elements of the GLDS vector.

To compute the GLDS vector $p_d(k)$ for a given mammographic ROI $H(i,j)$, and a given displacement vector $d = (d_1, d_2)$, first a difference image is computed as $H_d(i,j) = |H(i,j) - H(i+d_1, j+d_2)|$. The k^{th} entry of the vector p_d is defined as the probability of occurrence of the pixel value k in the difference image $H_d(i,j)$.

If the image texture is coarse, and the length of the displacement vector d is small compared to the texture element size, then the pixels separated by d will usually have similar pixel values. This implies that the elements of GLDS vector will be concentrated around 0, *i.e.*, $p_d(k)$ will be large for small values of k, and small for large values of k. Conversely, if the length of the vector d is comparable to the texture element size, then the elements of the GLDS vector will be distributed more evenly.

Since the image matrix is discrete, the displacement vector used in feature calculation is usually chosen to have a phase of $\theta = 0°$, $45°$, $90°$, or $135°$. These phases correspond to displacement vectors of $d = (d_0, 0)$, $d = (d_0, d_0)$, $d = (0, d_0)$, and $d = (d_0, -d_0)$, respectively. If image texture is directional, features computed at the same vector magnitude but different phases will convey useful and distinct information. In our case, we did not observe any directional preference in the texture-images that we calculated. Therefore, we averaged textures obtained at the same vector magnitude but different phases. The vector magnitudes at displacement vectors of $d = (d_0, 0)$, $d = (0, d_0)$ and $d = (d_0, d_0)$, $d = (d_0, -d_0)$ differ by a factor of $\sqrt{2}$. For this reason, we averaged the texture features obtained at $\theta = 0°$, $90°$ and $\theta = 45°$, $135°$ separately. To reduce the number of texture combinations, we used only the averages obtained at $\theta = 45°$, $135°$, *i.e.*, we averaged the texture features obtained at $d = (d_0, d_0)$ and $d = (d_0, -d_0)$. In the following discussion, we refer to this average as the feature obtained at a texture distance of d_0. The effect of different texture distances on classification was evaluated by studying the classification accuracy at texture distances of $d_0 = 2, 4$, and 8.

In this paper, we used four GLDS texture features, namely, contrast, angular second moment, entropy and mean, which are defined in the Appendix.

D.2 GLDS Texture-Images

Within a selected ROI, there might be several sub-regions showing different texture statistics, for example, the region inside the mass, the transition region between the mass and the surrounding tissue, and the surrounding tissue. If the texture is computed for the entire ROI, the computation result will be an average of the texture features for the different regions.

One can characterize these feature differences by computing the features in different sub-regions inside the ROI. In this study, we moved the center of the sub-region on a rectangular grid over the ROI, and considered each computed feature as the pixel value of a texture-image at that grid location. The texture-images were then input into a CNN for classification.

Each of the four GLDS features described in the Appendix, namely, contrast, angular second moment, entropy and mean, were used to obtain GLDS texture-images. To obtain a single pixel of a texture-image, one of these features was computed in an $R \times R$ sub-region of the ROI. To obtain pixel values of the texture-image at different pixel locations, the center of the sub-region was moved over the ROI on a rectangular grid with grid distance G. More precisely, the $(i,j)^{th}$ element of the texture-image was obtained from the $R \times R$ sub-region whose upper-left corner was at pixel location (Gi, Gj) in the original image. The computation of a texture-image is illustrated in Fig. 6.

The sub-regions might or might not overlap depending on the relation between G and R. The size of the texture-image was the smallest integer larger than $(M - R)/G$, where M was the original ROI size. The classification results with GLDS texture-images reported in Section IV.B. were obtained with $R = 30$ and $G = 15$.

D.3 CNN with GLDS Texture-Images

The CNN architecture employed for classifying mass and nonmass ROIs using GLDS texture-images is shown in Fig. 7. This CNN had a single hidden layer with three image groups, a single output, and two input images. The first input image was a 16×16 averaged-subsampled image that was also used alone for ROI classification with averaged-subsampled images. The second input image was a 16×16 texture-image obtained using one of four GLDS features, contrast, angular second moment, entropy, and mean.

As in the case of CNN with averaged-subsampled images, eight pairs of rotated and mirrored images belonging to each ROI were applied to the CNN consecutively. Since the GLDS features were calculated as the average of texture features at $\theta = 45°$ and $135°$, we did not have to re-calculate GLDS texture-images for the rotated or mirrored images. We only needed to rotate or mirror the GLDS texture-images, similar to the rotation and mirroring of the averaged-subsampled images. As in the case of CNN with averaged-subsampled images, forward and backpropagation were accomplished using Equations (3)-(8), this time with $N(1) = 2$, $N(2) = 3$, and $N(3) = 1$. Classification accuracy was evaluated using the same methods as in Section III.C.

E. Classification with SGLD Texture-Images

E.1 SGLD Features

A second method of defining statistical texture features is through the SGLD matrix. SGLD features were previously shown to be useful in distinguishing mass ROIs from normal tissue [11], [12]. To compute the SGLD matrix for an image $H(i,j)$, a displacement vector $d = (d_1, d_2)$ is defined. The $(k_1, k_2)^{th}$ element of the SGLD matrix, P_d, is defined as the joint probability that gray levels k_1 and k_2 occur at a distance of (d_1, d_2) in $H(i,j)$.

SGLD features are affected by the number of bits used to represent the image (bit depth). Images of lower bit depth can be derived from images of higher bit depth by eliminating the least significant bits. The choice of bit depth used in SGLD matrix computation is important because of the trade-off between the gray level resolution and the statistics of the estimated joint probability distribution. If the bit depth is high, then the number of pixels pairs that contribute to an element of the SGLD matrix will be low, and the statistics of the estimated joint probability distribution will be poor. The noise in the least significant bits of the image will also affect the distribution. On the other hand, if the bit depth is low, these two problems are alleviated, but some of the characteristic features of the distribution may be lost due to the reduced gray level resolution. Based on the results of [11], we used a bit depth of 7 bits in SGLD matrix construction.

SGLD features mainly reflect the distribution of the elements in the SGLD matrix. For example, the correlation measure defined in [20] is high when the entries are higher along the main diagonal of the SGLD matrix, and the entropy measure attains its maximum value when all the elements of the SGLD matrix are equal.

As in the case of GLDS features, we used displacement vectors with phases of $\theta = 45°$ and $135°$ for SGLD texture feature calculation. These phases corresponded to displacement vectors of $d = (d_0, d_0)$ and $d = (d_0, -d_0)$. Texture features obtained for displacement vectors of $d = (d_0, d_0)$ and $d = (d_0, -d_0)$ were averaged to obtain a GLDS feature at a texture distance of d_0. The effect of different texture distances on classification was evaluated by studying the classification accuracy at texture distances of $d_0 = 12, 16, 20$, and 24.

In this paper, we used three SGLD features, namely correlation, entropy, and difference entropy, which were among the best features for the classification of masses and benign tissue in a previous study [11]. The definitions of these features are given in the Appendix.

E.2 SGLD Texture-Images

The computation of SGLD texture-images parallels the computation of GLDS texture-images. Each of the three SGLD features described in the Appendix, namely correlation, entropy, and difference entropy, were used to obtain GLDS texture-images. To obtain a single pixel of a texture-image, one of these features was computed in an $R \times R$ sub-region of the ROI. To obtain pixel values of the texture-image at different pixel locations, the center of the sub-region was moved over the ROI on a rectangular grid with grid distance G. Fig. 6. describes the computation of texture-images pictorially. The classification results with SGLD texture-images reported in Section IV.C. were obtained with $R = 60$ and $G = 13$ for texture distances (d_0) of 12 and 16, and with $R = 75$ and $G = 12$ for texture distances of 20 and 24.

E.3 CNN with GLDS Texture-Images

The CNN architecture employed for classifying mass and nonmass ROIs using SGLD texture-images is the same as that used for classification with GLDS texture-images, and is shown in Fig. 7. This CNN had a single hidden layer with three image groups, a single output, and two input images. The first input image was a 16×16 averaged-subsampled image that was also used alone for ROI classification with averaged-subsampled images. The second input image was a 16×16 texture-image obtained using one of three SGLD features, correlation, entropy, and difference entropy. CNN training and performance evaluation was carried out similarly to Section III.C.

F. Classification with GLDS and SGLD Texture-Images

The CNN architecture employed for classifying mass and nonmass ROIs using both GLDS and SGLD texture-images is shown in Fig. 8. We investigated the use of a three-layer CNN with a three input images, and a single output node. The number of hidden layer groups $N(2)$, and the weight kernel size between the input layer and the hidden layer $S_w(1)$, were variable. The first input image was a 16×16 averaged-subsampled image, the second input image was a 16×16 texture-image obtained using the GLDS mean texture-image at a texture distance of $d_0 = 4$, and the third input image was a 16×16 texture-image obtained using the SGLD correlation texture-image at a texture distance of $d_0 = 16$. CNN training and performance evaluation was carried out similarly to Section III.C.

IV. Results

A. Results with subsampled images

For the purpose of computational efficiency, the image size was reduced by averaging adjacent pixels and subsampling, as described in Section III.C. The resulting 32×32 or 16×16 ROIs were used as inputs to a three-layer CNN with one image group at the input layer, $N(2)$ image groups at the hidden layer, and a single output node. We investigated the effect of varying the number of image groups $N(2)$, and the CNN weight kernel size $S_w(1)$ between the input layer and the hidden layer. The A_z values for the training and test sets are summarized in Table I for 16 × 16 input images, and in Table II for 32 × 32 input images, respectively. The training and test ROC curves for the CNN with 16 × 16 input images, $N(2) = 3$ and $S_w(1) = 10$ are plotted in Fig. 9, and the corresponding learning curves are plotted in Fig. 10.

B. Results with GLDS features

The results of Subsection IV.A indicate that the performance was not significantly different (i) between 16 × 16 and 32 × 32 input images; and (ii) among CNN architectures with different values of $N(2)$ and $S_w(1)$. For this reason, in this subsection, we chose to fix these variables while we studied the effect of the texture feature and distance variables.

All the CNNs in this subsection had two 16 × 16 input images, $N(2) = 3$, and $S_w(1) = 10$. The first input image was a 16 × 16 averaged-subsampled image that was also used in the previous subsection. The second input image was a 16 × 16 texture-image obtained using one of four GLDS features, contrast, angular second moment, entropy and mean. Training and test results are summarized in Table III for texture distances of $d_0 = 2, 4$, and 8.

C. Results with SGLD features

As in Subsection IV.B, the CNNs in this subsection had two 16 × 16 input images, $N(2) = 3$, and $S_w(1) = 10$. The first input image was an averaged-subsampled image, and the second input image was a texture-image obtained using one of three SGLD features, correlation, entropy and difference entropy. We used texture distances of $d_0 = 12$, 16, 20 and 24, because the study in [11] indicated that the best classification accuracy was obtained within this range. Training and test results are summarized in Table IV.

D. Results with GLDS and SGLD features

In Subsections IV.B and IV.C, the CNN architecture was kept fixed as we studied the effect of the texture feature and distance variables for GLDS and SGLD features. In this subsection, we chose one GLDS and one SGLD feature, and studied the effect of the CNN architecture as we did in Section IV.A, but in this case with three input images instead of a single input image. The first input image was a 16 × 16 averaged-subsampled image that was also used in the previous three subsections. The second image was a GLDS mean texture-image at $d_0 = 4$, and the third was an SGLD correlation texture-image at $d_0 = 16$. These texture-images were chosen because they seemed to yield better classification results than the other texture-images as shown in Tables III and IV. Examples of these three CNN input images, for a mass and three nonmass ROIs extracted from the same mammogram, are shown in Fig. 11, along with the background-corrected ROIs. We investigated the effect of varying $N(2)$, and $S_w(1)$. The A_z values for the training and test sets are summarized in Table V. The training and test ROC curves for the CNN architecture with $N(2) = 8$ and $S_w(1) = 10$ are plotted in Fig. 12, and the learning curves are plotted in Fig. 13.

V. Discussion

A comparison of Tables I and V reveals that texture-images significantly improve the classification performance. Considering rows with the same number of hidden-layer image groups and the same kernel size in Tables I and V, test A_z values in Table V are 0.04 to 0.06 higher than their counterparts in Table I. The best test A_z value in Table V reaches 0.87, which, as observed from Fig. 12, corresponds to a TPF of 90% at a FPF of 31%. Figs 10 and 13 indicate that as training continued beyond a certain epoch, test A_z fluctuated around a saturation value, and training A_z continued to increase. As the number of CNN input images was increased, we observed a decline in the CNN learning rate, *i.e.*, more training epochs were required for the the test A_z curve (bold lines in Fig. 10 and Fig. 13) to reach its maximum.

A comparison of different rows in Table I or Table V indicates that the effect of the CNN architecture on classification accuracy is less important than that of the use of texture-images. For example, in Table V, when the kernel size was fixed at 10, and the number of image groups was varied, the test A_z value did not change from its best value of 0.87. Test A_z values within Table I and Table V differed by 0.01 to 0.03 when the number of image groups was varied between 3 and 8, and the kernel size was varied between 8 and 12. When we varied the CNN architecture, we did not observe a significant change in the number of training epochs necessary for the test A_z curve to reach its maximum. The overall training time on a computer was longer when the kernel size and the number of image groups were large, since each training epoch took a longer time to run. One has to study all "reasonable" combinations of CNN architectures and texture feature variables in order to optimize the classification accuracy. However, since CNN training is computationally intensive, this would take an inordinate amount of time. Instead, we attempted to find the "best" combination of features, texture distance, and CNN architecture in two stages, within the constraint of computation time. First, in Sections IV.B and IV.C, we determined which features and texture distances yielded better classification results using a single CNN architecture. Then, in Section IV.D, we varied the CNN architecture while the features and texture distances were fixed. Clearly, this results in a "suboptimal" combination, which, nevertheless, produced satisfactory classification results. It may be possible to improve our results using CNNs that employ more than three input images, more than a single hidden layer, or more than a single output node. It may also be possible to use different techniques to derive different CNN input images from an ROI to further improve the classification accuracy. However, the results of our limited-scale study demonstrate the viability of our approach.

Since a neural network uses an iterative minimization technique in training, the initial state of a CNN is potentially important for training. To obtain an indication about the dependence of CNN performance on initial weight values, we initialized the CNN in the last row of Table V ($N(2) = 8$ and $S_w(1) = 10$) with five different seeds for the random number generator, which produced five different sets of initial weights. After training and testing, we computed the average and standard deviation of the test A_z obtained using these five sets of initial weights. The average A_z was 0.87, *i.e.* unchanged from the value in Table V, and the standard deviation was 0.002. This indicates that the performance of the CNN that we implemented is consistent in spite of random variations in the initial weights.

Results of Section IV.A. indicate that there is no significant difference in classification accuracy between CNNs that operate on 16 × 16 and 32 × 32 subsampled ROIs. However, this does not mean that resolution of the ROI does not have any effect on the classification accuracy. It may well be possible that 32 × 32 subsampled ROIs still do not contain enough detail to improve the classification results. It may also be possible to significantly improve the classification results by applying larger ROIs with better resolution to the CNN. When the computing power becomes available, we will explore the effect of the ROI resolution on classification accuracy.

A shift-invariant neural network (SINN) that is similar to CNN was applied in [14] to detection of microcalcifications on mammograms. CNN and SINN differ mainly in that the test and training outputs of a SINN are images, as opposed to real numbers. The output images of a SINN are processed using thresholding and segmentation techniques before classification is performed. The advantage of SINN is that it yields a spatially-invariant output, *i.e.*, ignoring edge effects, if the input ROI is translated, the output is also translated by the same amount. The advantages of CNN in mass detection include (i) in training, one does not need to supply a desired image to the CNN that contains the pixel locations of the true mass, which may be difficult to obtain in many cases; and (ii) after testing, no image processing is required to perform classification: Only a single threshold is used to separate the two classes. In our current application, spatial invariance is not very critical since the masses are centered in the manually-extracted ROIs. We are currently developing algorithms which will center the mass in an ROI obtained by an automated detection and extraction program.

In our laboratory, we have previously investigated two other classification techniques using the same ROI set as in this paper [11], [12]. The classification method in [11] employs SGLD texture features obtained at fixed distances, and the method in [12] employs multiresolution texture analysis. The best test results obtained in this paper ($A_z = 0.873$) are better than the best results in [11] ($A_z = 0.823$), and comparable to the best results in [12] ($A_z = 0.859$).

Our interest in CNN as an alternative classifier stems from the fact that different classifiers are potentially better suited to classify different types of masses and normal tissue. It is not yet possible to predict which masses will be more correctly classified by a CNN classifier and which masses will be more correctly classified by multiresolution texture analysis. However, our experiments have shown that combining the outputs of these two classifiers improves classification accuracy. In [26], combining the results of a CNN that operates on averaged-subsampled images (as in Section IV.A.), and a classifier that operates on multiresolution texture features, we obtained an A_z value of 0.89 with the same ROI set as in this paper. A more complete analysis of the application of different classifiers to the problem of ROI classification will be published elsewhere.

The long-term objective of this research is to develop a CAD system which will provide a second opinion to the radiologist concerning the presence of lesions on a mammogram. This long-term objective can be divided into two more-easily manageable goals: (i) detection of suspicious ROIs on a mammogram, and (ii) elimination of suspicious, but normal ROIs from the ROIs detected in (i). This paper deals with this latter goal. When the research on these two goals are integrated, it will be possible to conduct observer studies to evaluate the improvement in radiologists performance when they are assisted by CAD. Although we have not attempted to compare the ROI classification accuracy reported in this paper to that of the radiologists, we suspect that radiologists will perform significantly better than a CNN in classifying the ROIs in this paper. However, the contribution of the CAD system will not be whether the CAD outperforms the radiologists, but rather how much CAD assists radiologists in detecting lesions that would otherwise be missed.

VI. Conclusion

We studied the application of a convolution neural network to classification of masses and normal ROIs. CNN input images were derived from the ROIs using (i) averaging and subsampling; (ii) GLDS feature extraction; and (iii) SGLD feature extraction. Using a three-layer CNN and three input images derived from each ROI, we obtained an average test A_z of 0.87, which corresponded to an average true-positive fraction of 90% at a false positive fraction of 31%. Our results indicated that the choice of CNN input images is more important than the choice of CNN architecture. Although classification performance needs to be further improved in order for the classifier to be useful in a clinical setting, our study indicates that a CNN can be trained to effectively classify masses and normal breast tissue on mammograms. We are currently investigating the effectiveness of the CNN classifier for differentiation of masses and normal ROIs obtained with an automatic extraction algorithm as a step towards a fully automated computer-aided diagnosis scheme.

Acknowledgment

The authors are grateful to Charles E. Metz, Ph.D., for the LABROC1 programs.

Appendix

A. GLDS Features

Given a GLDS vector $p_d(k)$ described in Section III.D, the GLDS texture features used in this paper are defined as follows [19], where K is the dimension of $p_d(k)$.

1. Contrast:

$$\mathrm{CON} = \sum_{k=0}^{K-1} k^2 p_d(k),$$

2. Angular Second Moment:

$$\mathrm{ASM} = \sum_{k=0}^{K-1} p_d(k)^2,$$

3. Entropy:

$$\mathrm{G\_ENT} = -\sum_{k=0}^{K-1} p_d(k) \log p_d(k),$$

4. Mean:

$$\mathrm{MEAN} = \frac{1}{K} \sum_{k=0}^{K-1} k p_d(k).$$

B. SGLD Features

Given an SGLD matrix $P_d(k_1, k_2)$ described in Section III.E, the SGLD texture features used in this paper are defined as follows [20], where K is the size of $P_d(k_1, k_2)$.

1. Entropy:

$$\mathrm{S\_ENT} = -\sum_{k_1=0}^{K-1} \sum_{k_2=0}^{K-1} P_d(k_1, k_2) \log P_d(k_1, k_2),$$

2. Difference Entropy:

$$\mathrm{DIF\_ENT} = -\sum_{k=0}^{K-1} P_{x-y}(k) \log P_{x-y}(k),$$

where

$$P_{x-y}(k) = \sum_{k_1=0}^{K-1} \sum_{k_2=0}^{K-1} P_d(k_1, k_2), \quad |k_1 - k_2| = k,$$

3. Correlation:

$$\mathrm{COR} = \frac{1}{\sigma_x \sigma_y} \sum_{k_1=0}^{K-1} \sum_{k_2=0}^{K-1} (k_1 - \mu_x)(k_2 - \mu_y) P_d(k_1, k_2),$$

where

$$\begin{aligned}
\mu_x &= \sum_{k_1=0}^{K-1} k_1 \sum_{k_2=0}^{K-1} P_d(k_1, k_2), \\
\sigma_x^2 &= \sum_{k_1=0}^{K-1} (k_1 - \mu_x)^2 \sum_{k_2=0}^{K-1} P_d(k_1, k_2), \\
\mu_y &= \sum_{k_2=0}^{K-1} k_2 \sum_{k_1=0}^{K-1} P_d(k_1, k_2), \\
\sigma_y^2 &= \sum_{k_2=0}^{K-1} (k_2 - \mu_y)^2 \sum_{k_1=0}^{K-1} P_d(k_1, k_2).
\end{aligned}$$

References

[1] C. C. Boring, T. S. Squires, T. Tong, and S. Montgomery, "Cancer statistics, 1994," *CA-A Cancer Journal for Clinicians*, vol. 44, pp. 7–26, 1994.

[2] H. C. Zuckerman, "The role of mammography in the diagnosis of breast cancer," in *Breast Cancer, Diagnosis and Treatment* (I. M. Ariel and J. B. Cleary, eds.), pp. 152–172, New York: McGraw-Hill, 1987.

[3] R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiology*, vol. 184, pp. 613–617, 1992.

[4] M. L. Giger, "Computer-aided diagnosis," in *Syllabus: A Categorical Course in Physics Technical Aspects of Breast Imaging* (A. G. Haus and M. J. Yaffe, eds.), pp. 257–270, Oak Brook, Illinois: RSNA Publications, 1992.

[5] H.-P. Chan, K. Doi, C. J. Vyborny, and et al, "Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis," *Investigative Radiology*, vol. 25, pp. 1102–1110, 1990.

[6] L. W. Bassett, D. H. Bunnell, R. H. Jahashahi, R. Gold, R. D. Arndt, and J. Linsman, "Breast cancer detection: one versus two views," *Radiology*, vol. 165, pp. 95–97, 1987.

[7] S. M. Lai, X. Li, and W. F. Bischof, "On techniques for detecting circumscribed masses in mammograms," *IEEE Transactions on Medical Imaging*, vol. 8, pp. 377–386, 1989.

[8] D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammography," *IEEE Transactions on Medical Imaging*, vol. 9, pp. 233–241, 1990.

[9] F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vybomy, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analyisis of bilateral subtraction images," *Medical Physics*, vol. 18, pp. 955–963, 1991.

[10] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," *Radiology*, vol. 191, pp. 331–337, 1994.

[11] H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," *Physics of Medicine and Biology*, vol. 40, pp. 857–876, 1995.

[12] D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," *Medical Physics*, vol. 22, pp. 1501–1513, 1995.

[13] Y. Wu, K. Doi, M. L. Giger, and R. M. Nishikawa, "Compterized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks," *Medical Physics*, vol. 19, pp. 555–560, 1992.

[14] W. Zhang, K. Doi, M. L. Giger, Y. Wu, and R. M. Nishikawa, "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant neural network," *Medical Physics*, vol. 21, pp. 517–524, 1994.

[15] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81–87, 1993.

[16] S.-C. B. Lo, M. T. Freedman, J. S. Lin, and S. K. Mun, "Automatic lung nodule detection using profile matching and

backpropagation neural network techniques," *Journal of Digital Imaging*, vol. 6, pp. 48–54. 1993.

[17] S.-C. B. Lo, J.-S. J. Lin. M. T. Freedman, and S. K. Mun. "Computer-assisted diagnosis of lung nodule detection using artificial convolution neural network." in *Proceedings of SPIE Medical Imaging: Image Processing*, vol. 1898, (Newport Beach. CA), pp. 859–869. June 1992.

[18] H.-P. Chan. S.-C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," *Medical Physics*, vol. 22, pp. 1555–1567, 1995.

[19] J. S. Weszka, C. R. Dyer, and A. Rosenfeld. "A comparative study of texture measures for terrain classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, pp. 269–285, 1976.

[20] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification." *IEEE Transactions on Systems, Man, and Cybernetics*. vol. 3, pp. 610–621, 1973.

[21] K. Fukushima, S. Miyake, and T. Ito, "Neocognitron: A neural network model for a mechanism of visual pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 13, pp. 826–834. 1983.

[22] R. A. Jacobs. "Increased rates of conversion through learning rate adaptation," *Neural Networks*, vol. 1, pp. 295–307, 1988.

[23] B. Widrow and M. A. Lehr. "30 years of adaptive neural networks: Perceptron, madaline, and backpropagation," *Proceedings of the IEEE*, vol. 78, pp. 1415–1442, 1990.

[24] C. E. Metz. "ROC methodology in radiographic imaging," *Investigative radiology*, vol. 21, pp. 720–733, 1986.

[25] C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binormal ROC curve from continuously distributed test results," presented at the 1990 Annual Meeting of the American Statistical Association, Anaheim. CA, Aug. 1990.

[26] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsit. "Image classification using artificial neural networks," in *Proceedings of SPIE Medical Imaging: Image Processing*, vol. 2434, (San Diego, CA), pp. 838–845, Mar. 1995.

TABLE I

CNN CLASSIFICATION PERFORMANCE WITH 16 × 16 SUBSAMPLED IMAGES.

Kernel size	Number of groups	Training A_z	Test A_z
6	4	0.82	0.80
8	4	0.82	0.81
10	4	0.85	0.81
12	4	0.87	0.82
14	4	0.87	0.81
10	3	0.87	0.83
10	6	0.85	0.82
10	8	0.83	0.81

TABLE II

CNN CLASSIFICATION PERFORMANCE WITH 32 × 32 SUBSAMPLED IMAGES.

Kernel size	Number of groups	Training A_z	Test A_z
11	4	0.81	0.80
16	4	0.84	0.80
20	4	0.84	0.83
23	4	0.84	0.82
20	3	0.84	0.82
20	6	0.84	0.82
20	8	0.84	0.82

TABLE III

CNN CLASSIFICATION PERFORMANCE WITH TWO INPUT IMAGES DERIVED FROM AN ROI. THE FIRST IMAGE IS THE AVERAGED AND SUBSAMPLED IMAGE, THE SECOND IMAGE IS THE GLDS TEXTURE-IMAGE. ASM, CON, G_ENT, AND MEAN STAND FOR ANGULAR SECOND MOMENT, CONTRAST, ENTROPY, AND MEAN, RESPECTIVELY.

	$d_0 = 2$		$d_0 = 4$		$d_0 = 8$	
Feature	Training A_z	Test A_z	Training A_z	Test A_z	Training A_z	Test A_z
ASM	0.84	0.82	0.87	0.82	0.86	0.82
CON	0.84	0.82	0.91	0.82	0.86	0.83
G_ENT	0.90	0.84	0.91	0.85	0.89	0.84
MEAN	0.90	0.84	0.90	0.86	0.88	0.85

TABLE IV

CNN CLASSIFICATION PERFORMANCE WITH TWO INPUT IMAGES DERIVED FROM AN ROI. THE FIRST IMAGE IS THE AVERAGED AND SUBSAMPLED IMAGE, THE SECOND IMAGE IS THE SGLD TEXTURE-IMAGE. COR, DIF_ENT, AND S_ENT STAND FOR CORRELATION, DIFFERENCE ENTROPY, AND ENTROPY, RESPECTIVELY.

	$d_0 = 12$		$d_0 = 16$		$d_0 = 20$		$d_0 = 24$	
Feature	Training A_z	Test A_z	Training A_z	Test A_z	Training A_z	Test A_z	Training A_z	Test A_z
COR	0.87	0.84	0.84	0.84	0.85	0.83	0.85	0.81
DIF_ENT	0.86	0.82	0.86	0.84	0.86	0.83	0.85	0.82
S_ENT	0.84	0.84	0.86	0.84	0.86	0.83	0.85	0.83

TABLE V

CNN CLASSIFICATION PERFORMANCE WITH THREE INPUT IMAGES DERIVED FROM AN ROI. THE FIRST IMAGE IS THE AVERAGED AND SUBSAMPLED IMAGE, THE SECOND IMAGE IS THE GLDS MEAN TEXTURE-IMAGE AT $d_0 = 4$, THE THIRD IMAGE IS THE SGLD CORRELATION TEXTURE-IMAGE AT $d_0 = 16$.

Kernel size	Number of groups	Training A_z	Test A_z
6	3	0.84	0.83
8	3	0.90	0.84
10	3	0.90	0.87
12	3	0.91	0.86
10	4	0.89	0.87
10	6	0.89	0.87
10	8	0.91	0.87

Figure Captions

Fig. 1. Basic CNN architecture.

Fig. 2. An example of the mass and and normal ROIs selected from one of the mammograms used in this study. The four ROIs are: upper left–mass; upper right–mixed dense/fatty tissue; lower left–dense tissue; lower right–fatty tissue.

Fig. 3. The averaging boxes and distances used in background correction.

Fig. 4. An example of background correction. (a) Original ROI that contains a malignant mass. (b) Background corrected ROI.

Fig. 5. The CNN architecture used for ROI classification with averaged-subsampled ROIs.

Fig. 6. Computation of texture-images.

Fig. 7. The CNN architecture used for ROI classification with averaged-subsampled ROIs plus a texture-image.

Fig. 8. The CNN architecture used for ROI classification with averaged-subsampled ROIs plus the GLDS mean texture-image, and the SGLD correlation texture-image.

Fig. 9. ROC curve for CNN with the 16×16 averaged-subsampled input image, $N(2) = 3$, and $S_w(1) = 10$. The A_z value was 0.87 for training and 0.83 for test.

Fig. 10. Training and test A_z values versus training epoch number for the CNN in Fig. 9.

Fig. 11. Background corrected image, subsampled image, GLDS mean texture-image at $d_0 = 4$, and SGLD correlation texture-image at $d_0 = 16$ for (a) a mass ROI, as shown in Fig., 4b., and (b-d) three nonmass ROIs extracted from the same mammogram.

Fig. 12. ROC curve for CNN with three input images, $N(2) = 8$ and $S_w(1) = 10$. The first input image is the 16×16 averaged-subsampled image, the second image is the GLDS mean texture-image at $d_0 = 4$, and the third image is the SGLD correlation texture-image at $d_0 = 16$. The A_z value was 0.91 for training and 0.87 for test.

Fig. 13. Training and test A_z values versus training epoch number for the CNN in Fig. 12.

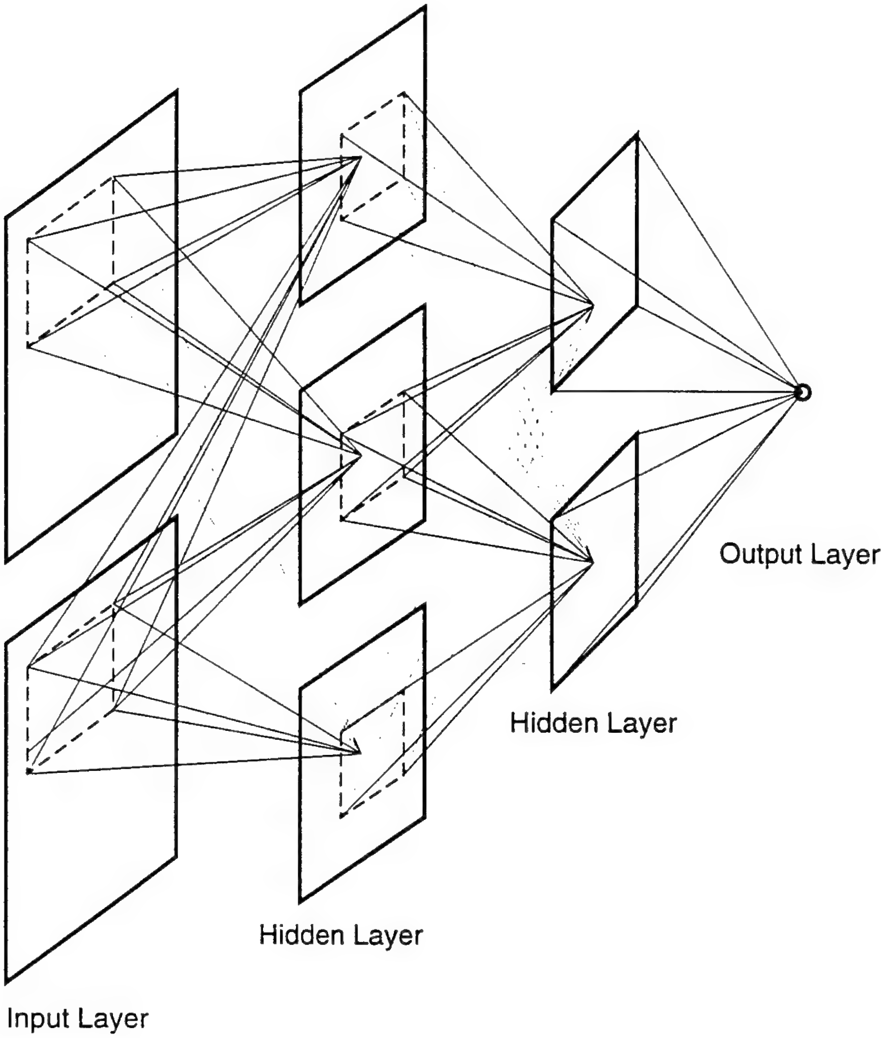


Fig. 1.

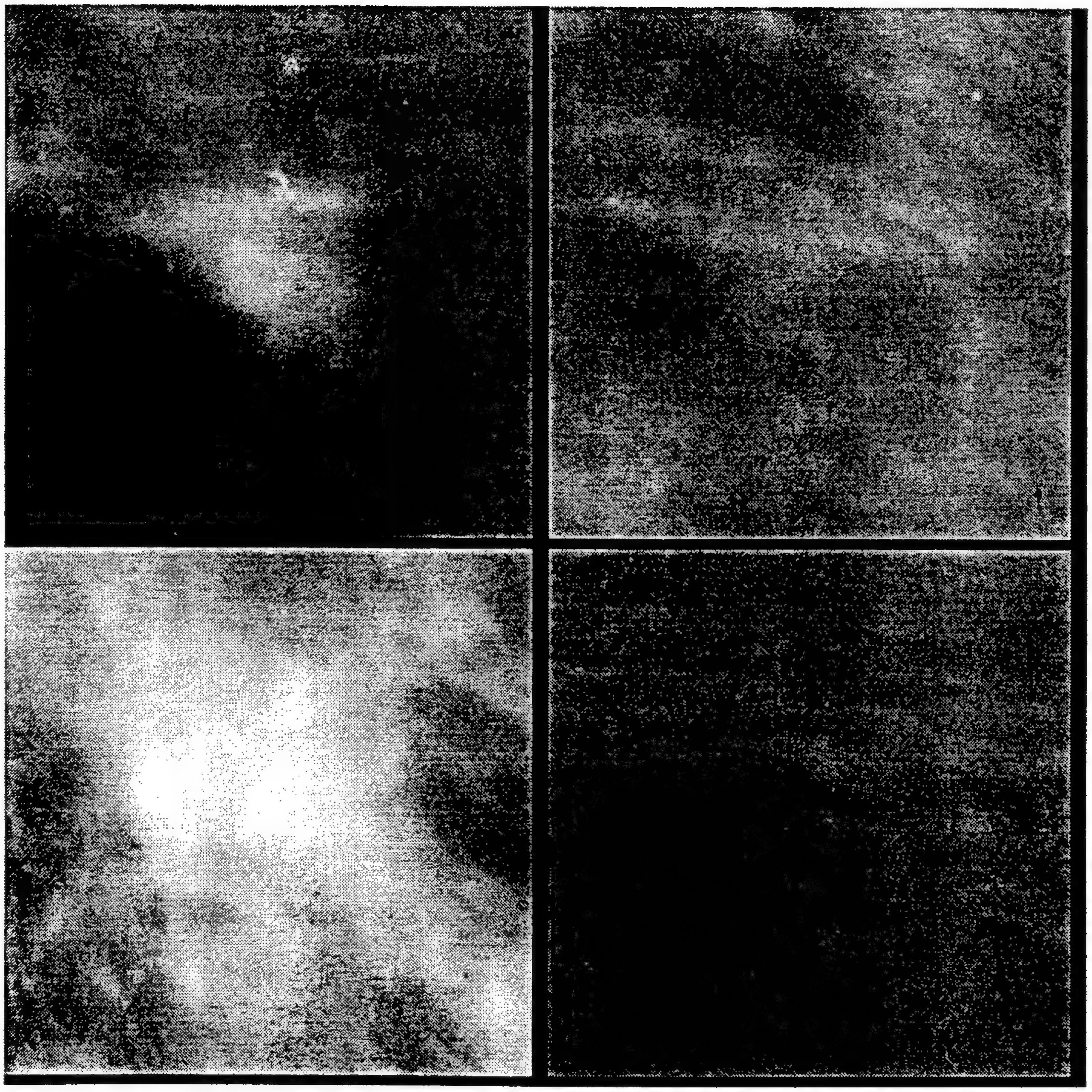


Fig. 2.

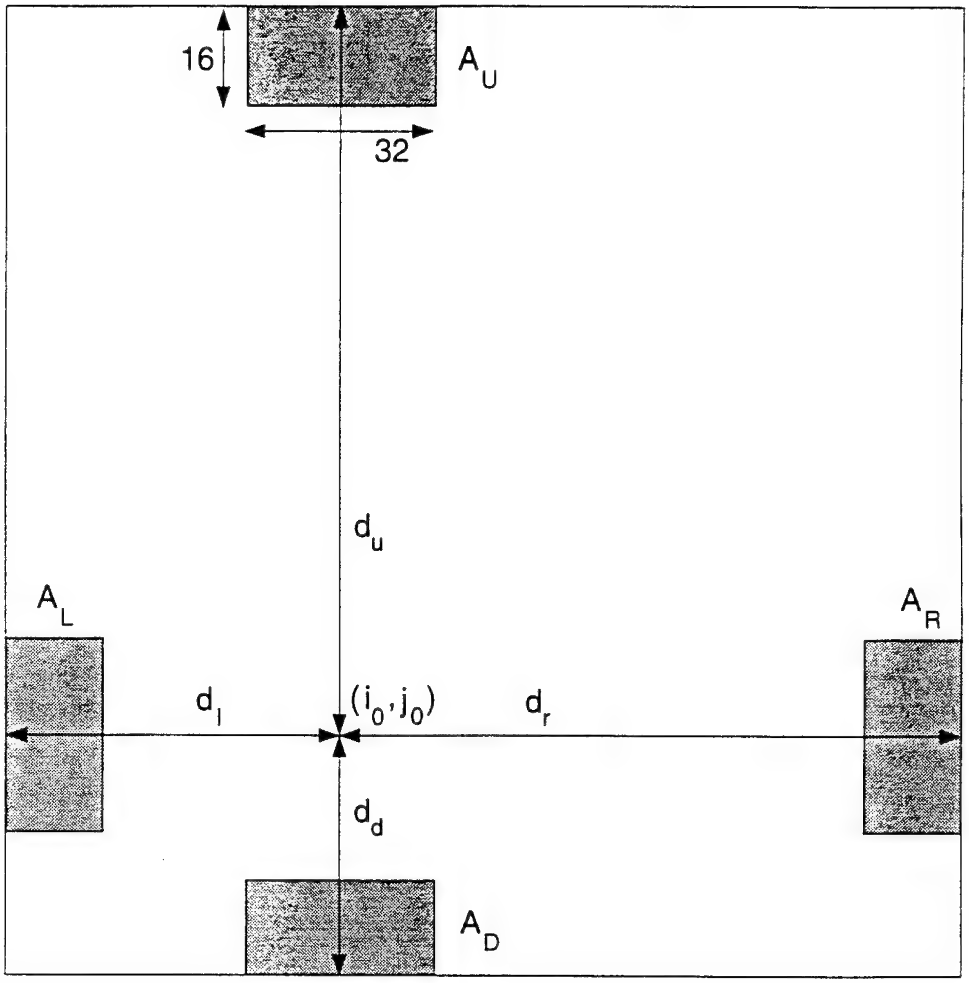


Fig. 3.

Fig. 4a.

Fig. 4b.

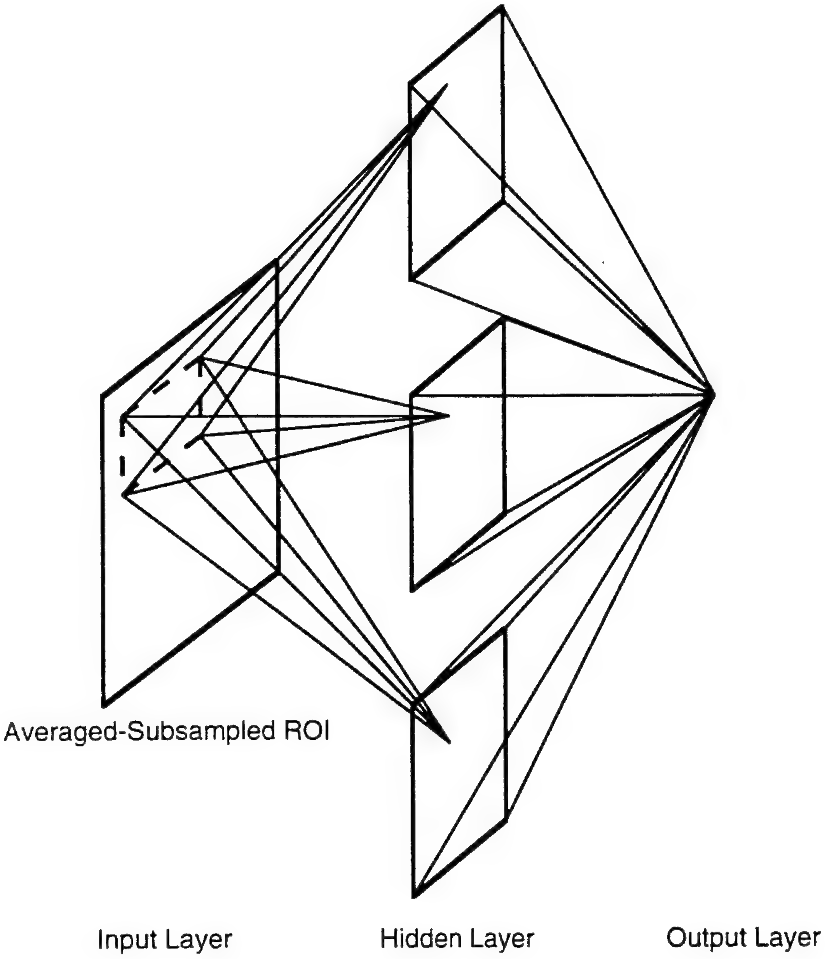

Averaged-Subsampled ROI
Input Layer
Hidden Layer
Output Layer

Fig. 5.

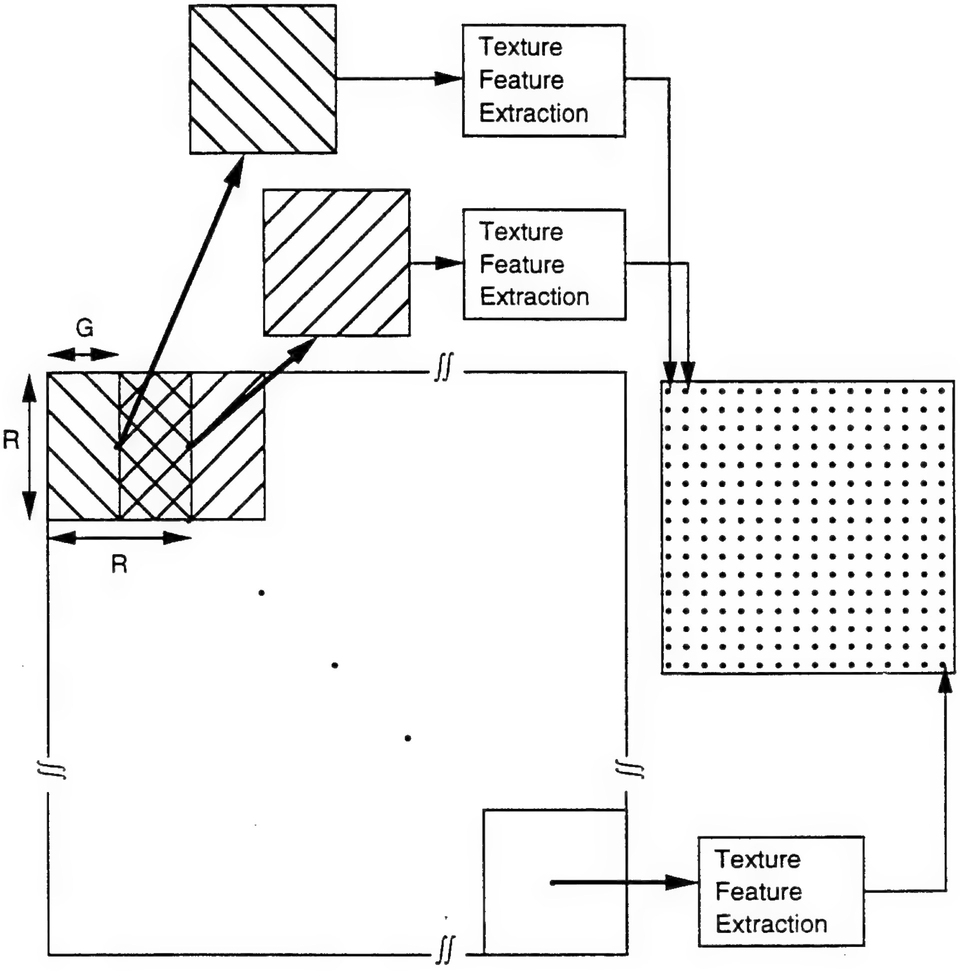

Texture
Feature
Extraction
Texture
Feature
Extraction
G
R
R
Texture
Feature
Extraction

Fig. 6.

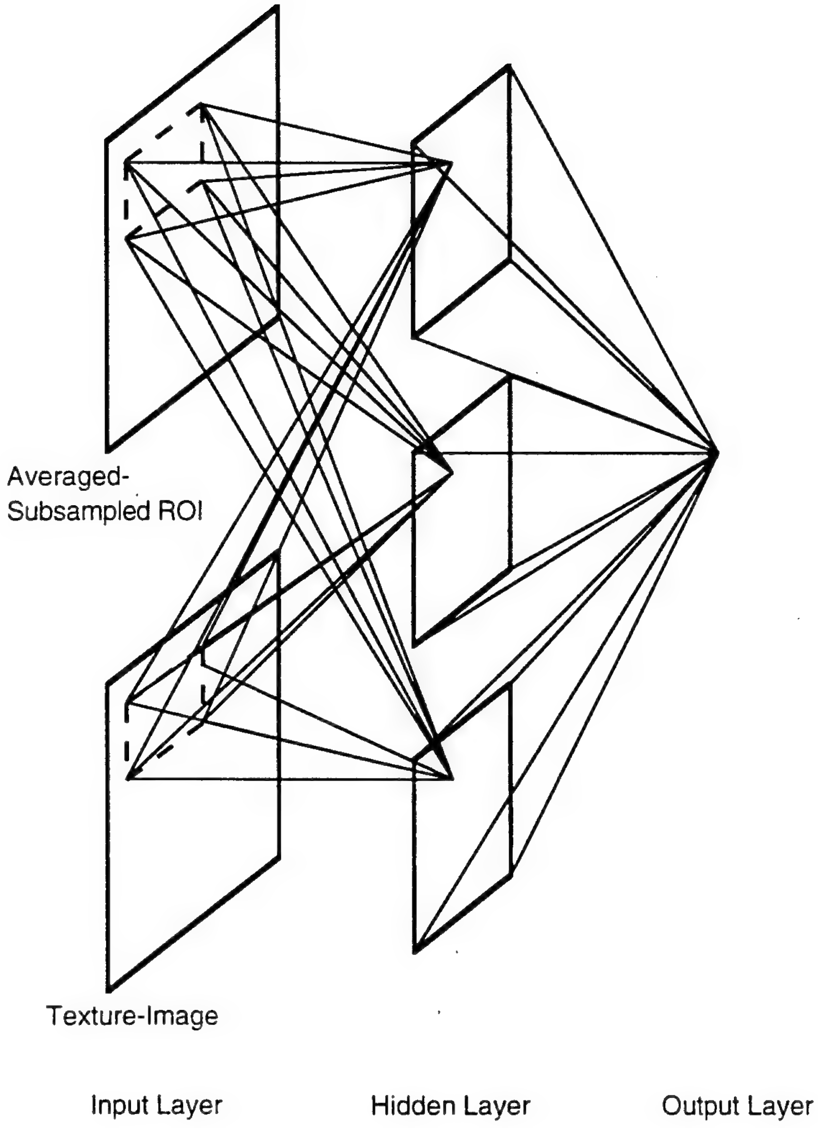

Averaged-
Subsampled ROI
Texture-Image
Input Layer
Hidden Layer
Output Layer

Fig. 7.

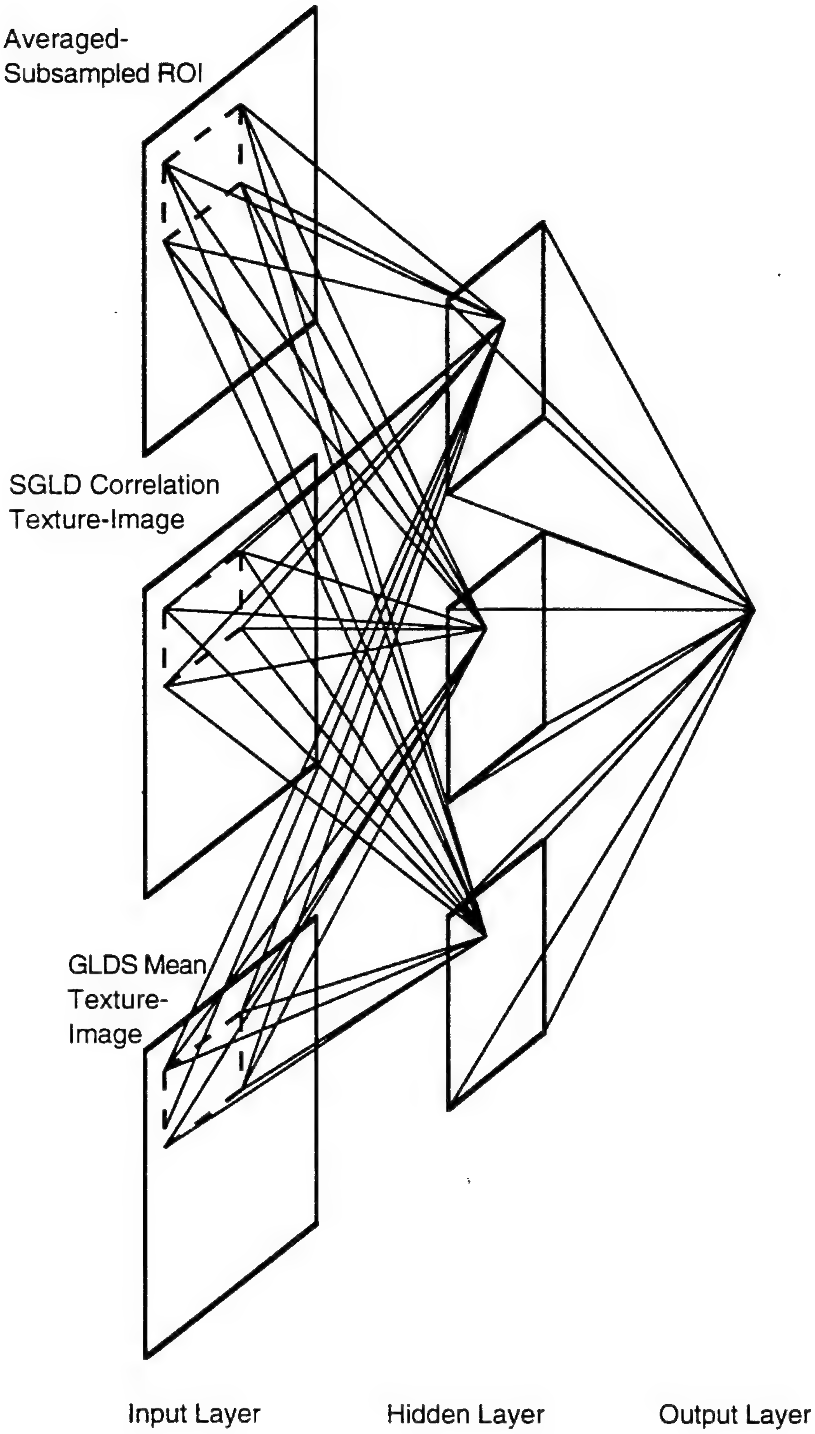

Averaged-
Subsampled ROI
SGLD Correlation
Texture-Image
GLDS Mean
Texture-
Image
Input Layer
Hidden Layer
Output Layer

Fig. 8.

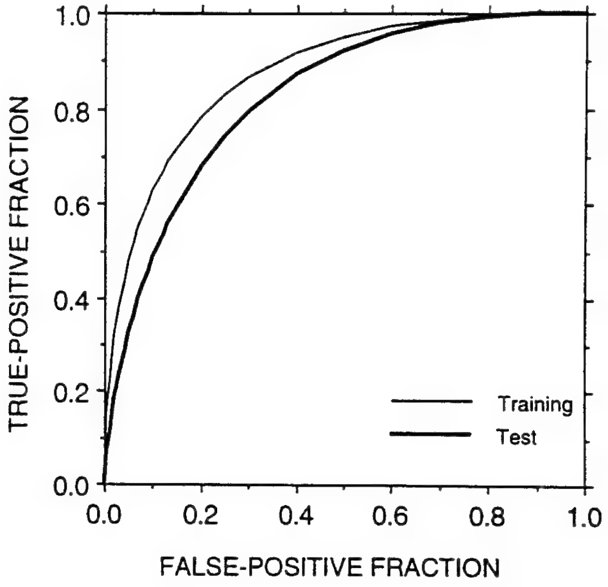

1.0
0.8
0.6
0.4
0.2
0.0
TRUE-POSITIVE FRACTION
0.0
0.2
0.4
0.6
0.8
1.0
FALSE-POSITIVE FRACTION
Training
Test

Fig. 9.

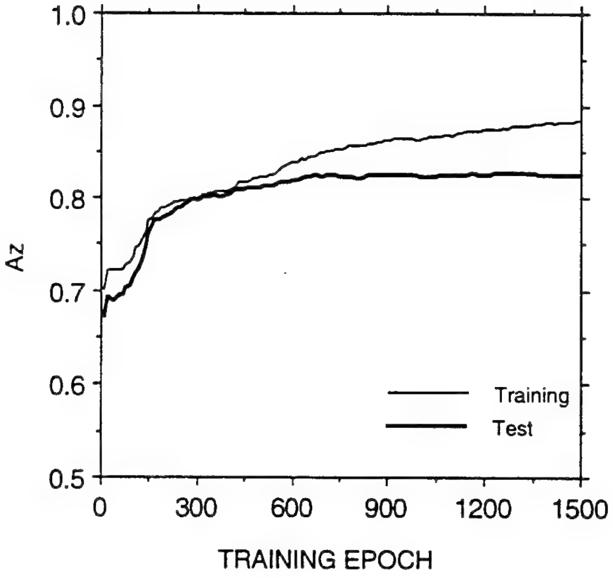

1.0
0.9
0.8
0.7
0.6
0.5
Az
0
300
600
900
1200
1500
TRAINING EPOCH
Training
Test

Fig. 10.

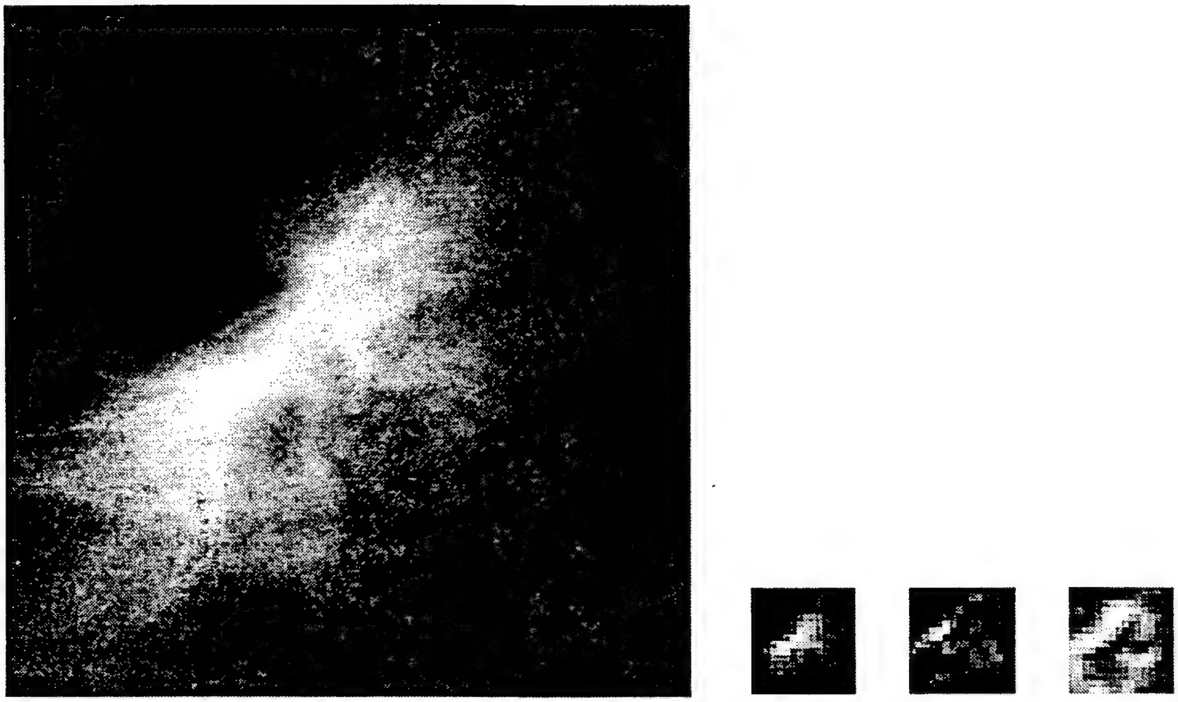


Fig. 11a.

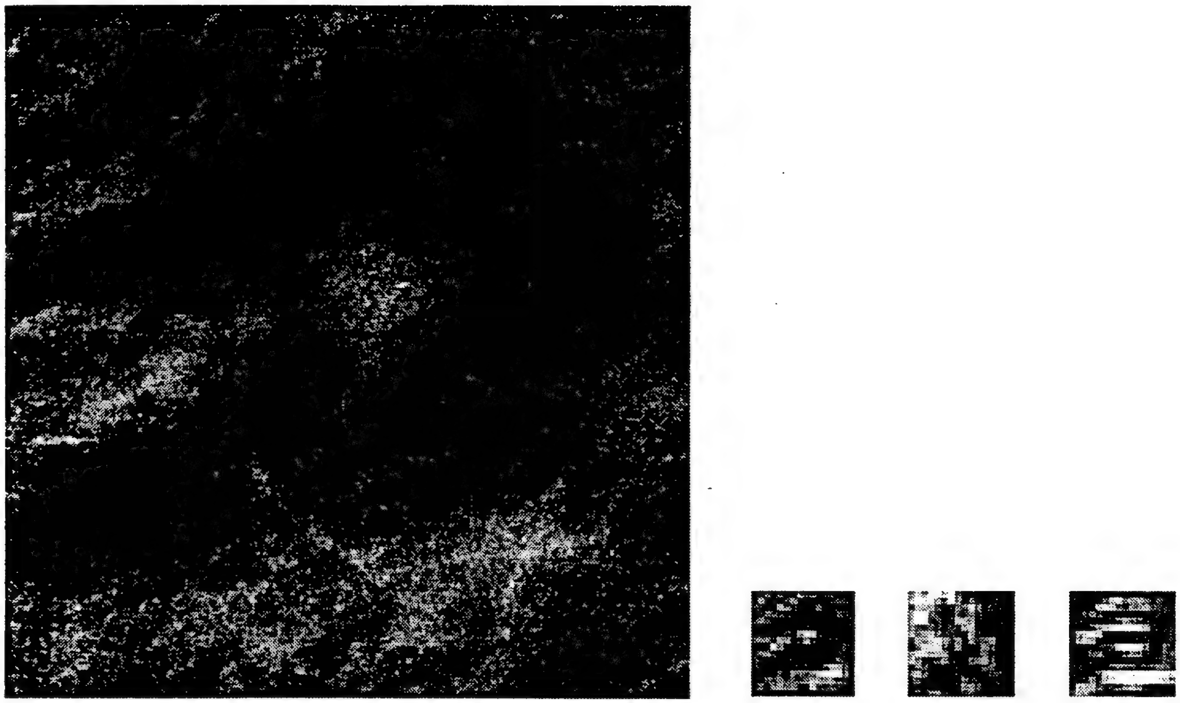


Fig. 11b.

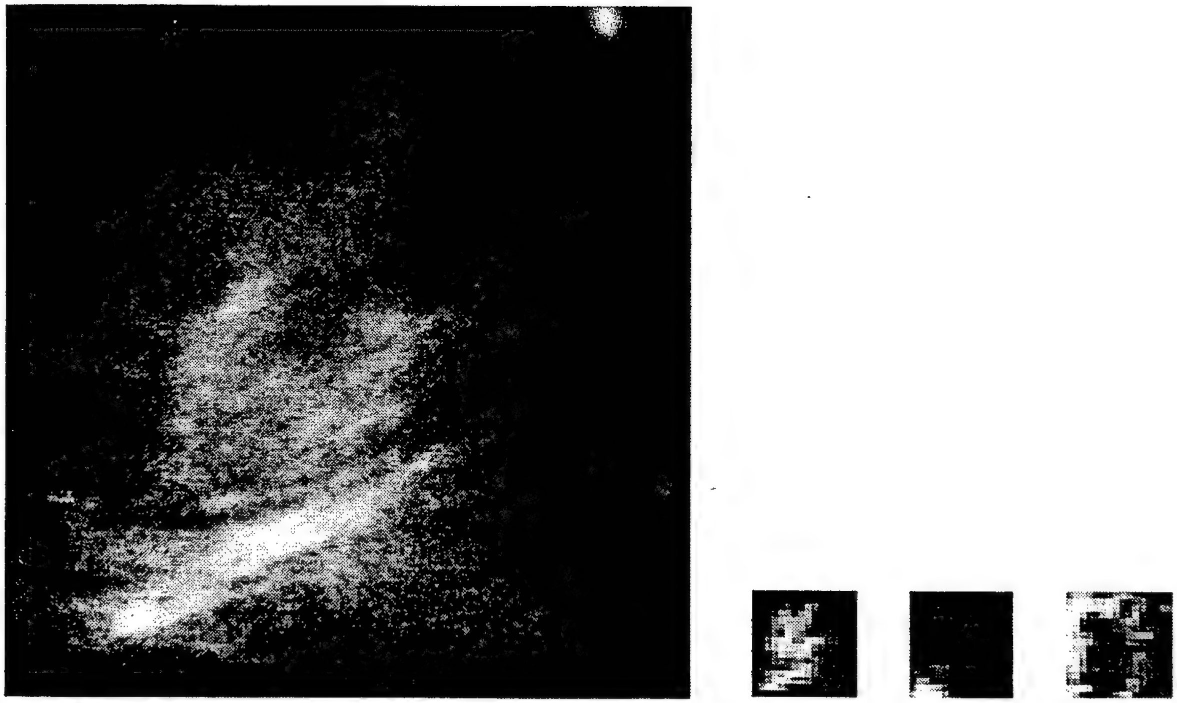


Fig. 11c.

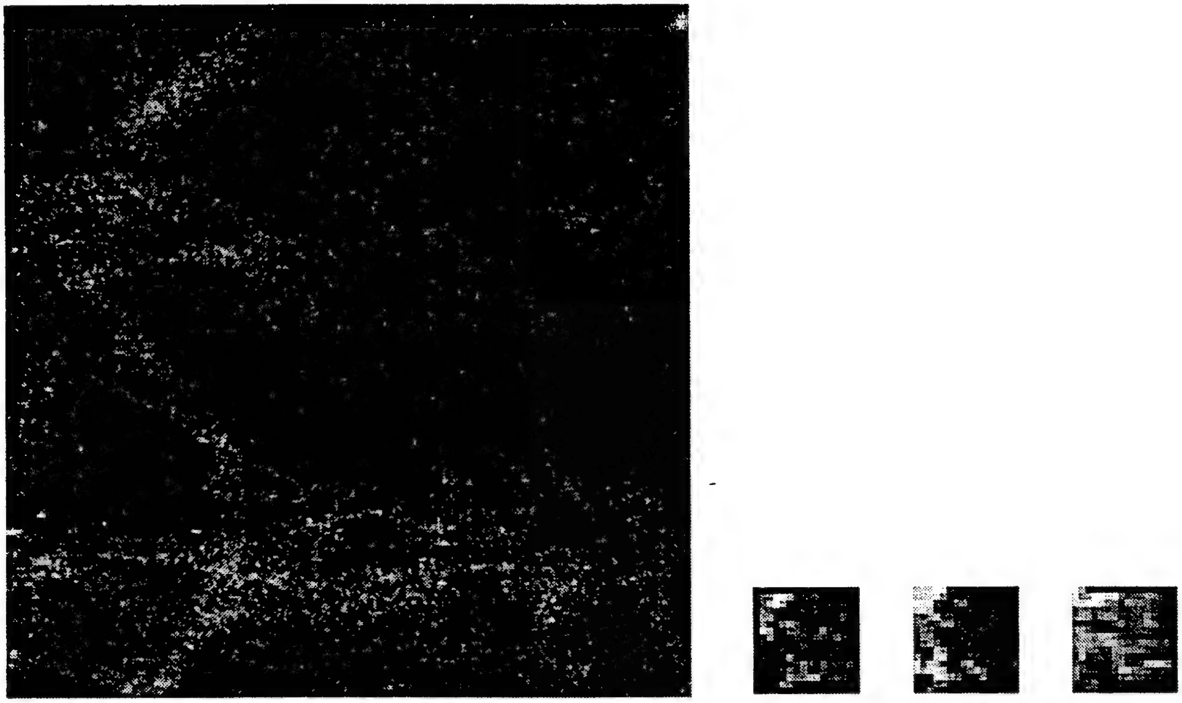


Fig. 11d.

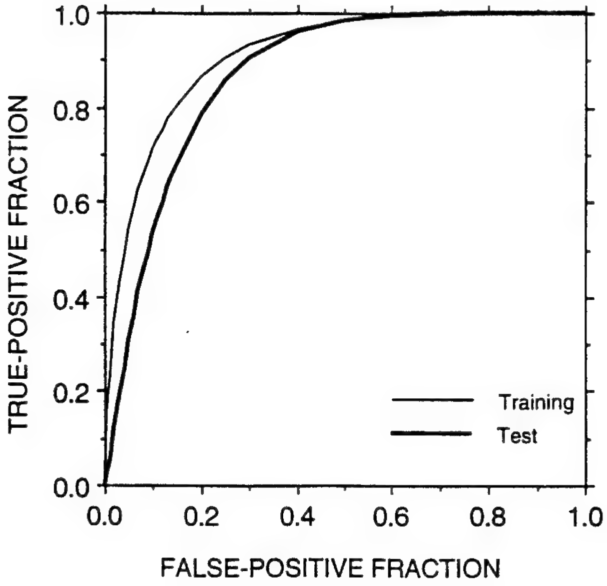

1.0
0.8
0.6
0.4
0.2
0.0
TRUE-POSITIVE FRACTION
0.0
0.2
0.4
0.6
0.8
1.0
FALSE-POSITIVE FRACTION
Training
Test

Fig. 12.

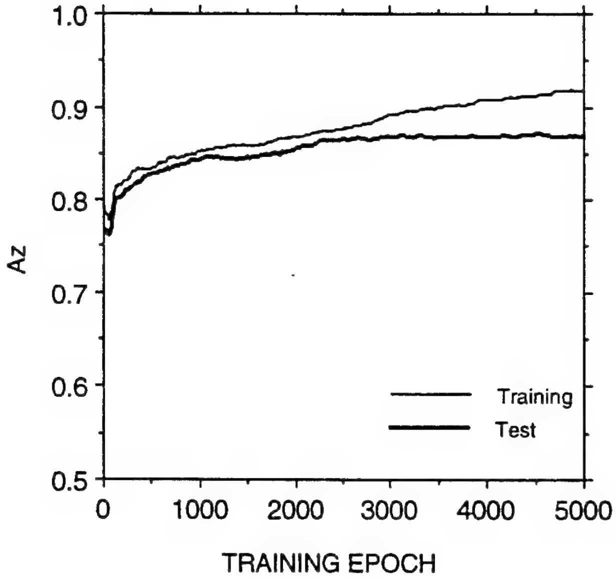

1.0
0.9
0.8
0.7
0.6
0.5
Az
0
1000
2000
3000
4000
5000
TRAINING EPOCH
Training
Test

Fig. 13.

Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis

Datong Wei, Heang-Ping Chan,[a)] Mark A. Helvie, Berkman Sahiner, Nicholas Petrick, Dorit D. Adler, and Mitchell M. Goodsitt
Department of Radiology, University of Michigan, Ann Arbor, Michigan

(Received 16 August 1994; accepted for publication 31 May 1995)

We investigated the feasibility of using multiresolution texture analysis for differentiation of masses from normal breast tissue on mammograms. The wavelet transform was used to decompose regions of interest (ROIs) on digitized mammograms into several scales. Multiresolution texture features were calculated from the spatial gray level dependence matrices of (1) the original images at variable distances between the pixel pairs, (2) the wavelet coefficients at different scales, and (3) the wavelet coefficients up to certain scale and then at variable distances between the pixel pairs. In this study, 168 ROIs containing biopsy-proven masses and 504 ROIs containing normal parenchyma were used as the data set. The mass ROIs were randomly and equally divided into training and test groups along with corresponding normal ROIs from the same film. Stepwise linear discriminant analysis was used to select optimal features from the multiresolution texture feature space to maximize the separation of mass and normal tissue for all ROIs. We found that texture features at large pixel distances are important for the classification task. The wavelet transform can effectively condense the image information into its coefficients. With texture features based on the wavelet coefficients and variable distances, the area A_z under the receiver operating characteristic curve reached 0.89 and 0.86 for the training and test groups, respectively. The results demonstrate that a linear discriminant classifier using the multiresolution texture features can effectively classify masses from normal tissue on mammograms.

Key words: mammography, computer-aided diagnosis, mass, wavelet transform, multiresolution texture analysis, linear discriminant classifier

I. INTRODUCTION

Mammography is considered the most reliable method for the early detection of breast cancers.[1] However, it has been reported that radiologists do not detect all breast cancers that are visible on mammograms in retrospective studies.[2–5] Previous studies indicate that computer-aided diagnosis (CAD) can provide a second opinion to the radiologists and potentially decrease the missed detection rate.[6,7] Computerized classification of the malignant or benign features of an abnormality may also be expected to reduce the number of negative biopsies. Improvement in the accuracy of mammography will increase its efficacy for screening and diagnosis of breast cancer.

Computer vision and artificial intelligence techniques have been developed to detect or characterize abnormalities on digital mammograms.[8] Image processing is usually a first step in computer vision to enhance the signal-to-noise characteristics of the objects being detected. Features are then extracted for classification between the signal and the background. Microcalcifications are ideal targets for computer detection due to their clinical relevance, their potential subtlety, and the lack of coexisting normal structures that have the same appearance.[8] The detection and classification of microcalcifications have received a lot of attention and demonstrated significant progress. Breast masses are more difficult to detect and classify than microcalcifications because masses can be simulated or obscured by normal breast parenchyma.[9,10] Fourier analysis of the masses does not show consistent and distinctive high-frequency components. Most of the signal (mass) energy is in the low-frequency region and overlaps with the frequency components of the normal tissue. The gray level changes at the mass boundary are usually gradual and not as abrupt as those at the boundary of microcalcifications. Moreover, the shape, size, and the gray level profile of the masses vary from case to case. These cause difficulties in the application of conventional image processing methods to the detection and feature characterization of masses.

Morphological features have been used to decrease the number of false-positive detections.[11] Spiculated masses were the focus of detection in the analysis of edge orientation in Kegelmeyer's work.[7,12] Breast cancers can also manifest as circumscribed masses.[13,14] Selective median filtering and template matching techniques were proposed to detect suspicious circumscribed masses.[14] For both types of masses, texture features were extracted from regions of interest (ROIs) in digital mammograms and were used in a decision tree to classify the masses from normal tissue with some success.[15]

The discovery of cortical neurons which respond specifically to stimuli within certain orientations and spatial frequencies suggests that multiorientation and multiresolution are part of the biological mechanism of the human visual system.[16,17] Interest in multiresolution image analysis has been growing rapidly in the field of computer vision. A multiresolution representation provides a simple hierarchical framework for analyzing image information. The compression of images by wavelet transforms can achieve a high compression ratio without significant loss of image details,[18] indicating that important image features are condensed in the wavelet coefficients. Texture analysis in the wavelet trans-

0094-2405/95/22(9)/1501/13/$6.00

form domain was used to distinguish different texture patterns (e.g., French canvas, beach sand, and oriental straw cloth) with some success.[19] Wavelet transform has been applied to mammographic image processing, especially to the enhancement and detection of microcalcifications. Laine *et al.*[17,20] proposed adaptive multiscale processing with wavelet decomposition and reconstruction for feature analysis and contrast enhancement. Richardson[21] discussed the use of wavelet packets that can be superior to wavelets for certain classes of mammographic signals. Qian *et al.*[22] proposed a tree-structured nonlinear adaptive filter and the wavelet transform for the detection and segmentation of microcalcifications on mammograms.

In this paper, we discuss the application of multiresolution texture analysis to digitized mammograms to distinguish mass from normal tissue. Multiresolution texture features were extracted from the spatial gray level dependence (SGLD) matrices (1) of the original image at variable distances, (2) of the wavelet coefficients at different scales, and (3) of the wavelet coefficients up to certain scales and then at variable distances, forming three feature vectors for each ROI. We used stepwise linear discriminant analysis to select features from each of these three texture spaces to maximize the separation of masses and normal tissue. The ability of the three feature vectors for classifying mammographic masses and normal tissue was compared. Receiver operating characteristic analysis was used to evaluate the classification accuracy of the texture features from the different feature spaces.

II. METHODS

A. Database selection

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsies in the Department of Radiology at the University of Michigan. The mammograms were acquired with dedicated mammographic systems with a 0.3-mm focal spot, a molybdenum anode, 0.03-mm-thick molybdenum filter, and a 5:1 reciprocating grid or a stationary grid. The image receptor was a Kodak MinR/MRE screen/film system with extended cycle processing. Our selection criterion was that a biopsy-proven mass could be seen on the mammogram. Initially, more than 300 mammograms were acquired. To avoid the effect of the repetitive grid pattern on the texture feature calculation and the classification, all mammograms with grid lines were excluded. Our final data set was composed of 168 mammograms.

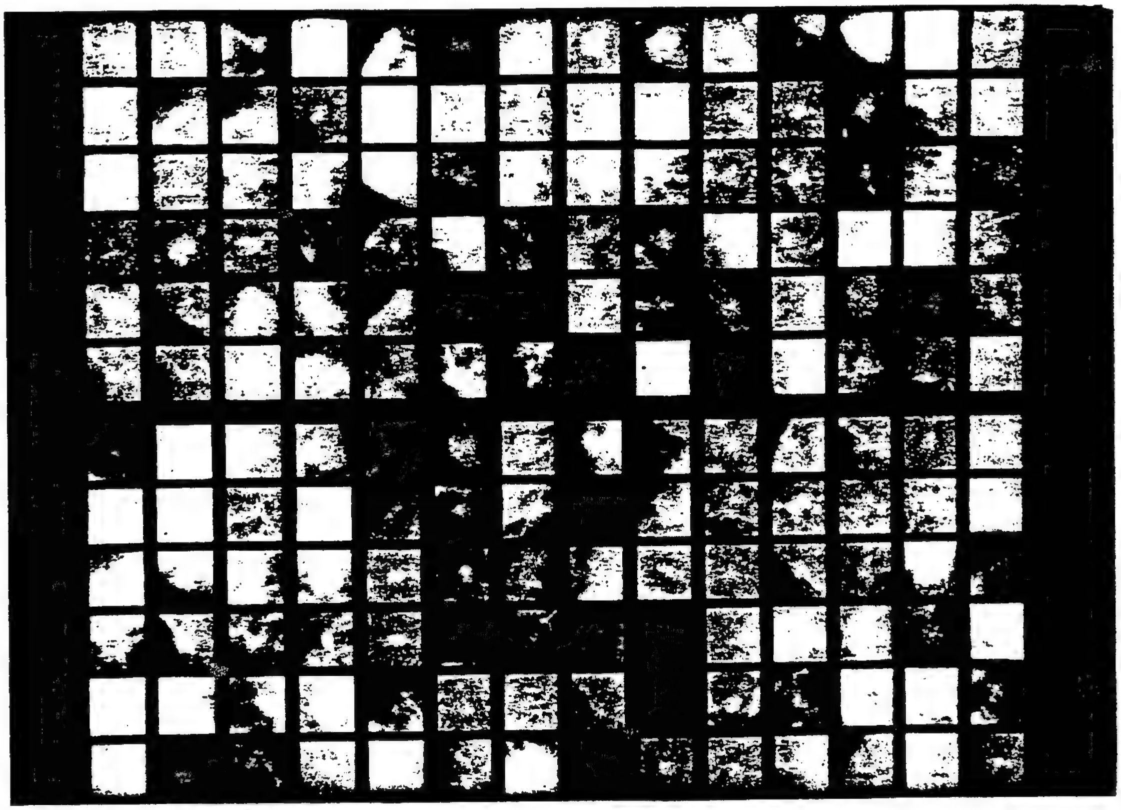
The mammograms were digitized with a laser film scanner (LUMISYS DIS-1000) at a pixel size of 0.1 mm×0.1 mm and 4096 gray levels. The light transmitted through the mammographic films was amplified logarithmically before digitization. After the calibration, the pixel values were linearly proportional to the optical density in the range of 0.1–2.8 optical density units. The slope of the calibration curve decreases gradually outside this range.

Before an automated computer segmentation procedure was developed, we used manual ROI extraction to study the feasibility of using texture features for the classification of mass and normal tissue in all types of breast parenchyma. Four different ROIs, each with 256×256 pixels, were selected manually from each mammogram. One ROI contained a true mass which was identified by an experienced mammographer. A second contained normal parenchyma including the densest tissue on that mammogram, a third, mixed dense/fatty tissue, and a fourth, fatty tissue. Figure 1 shows the 672 ROIs from the 168 mammograms in reduced spatial resolution. The 168 case samples in the data set contained a mixture of benign ($n=83$) and malignant ($n=85$) masses. Forty-five of the malignant masses and six of the benign masses were spiculated. The visibility of the masses was ranked by experienced radiologists on a scale of 1–10 (1 =most obvious, 10=most subtle), which corresponded to the range of masses seen on clinical mammograms. The length of the long axis (size) of the masses was also measured by the radiologists. The distributions of the visibility scores and the sizes are shown in Fig. 2. It can be seen from Figs. 1 and 2 that the masses with different shapes and visibility found in clinical practice were fairly well represented in the data set.

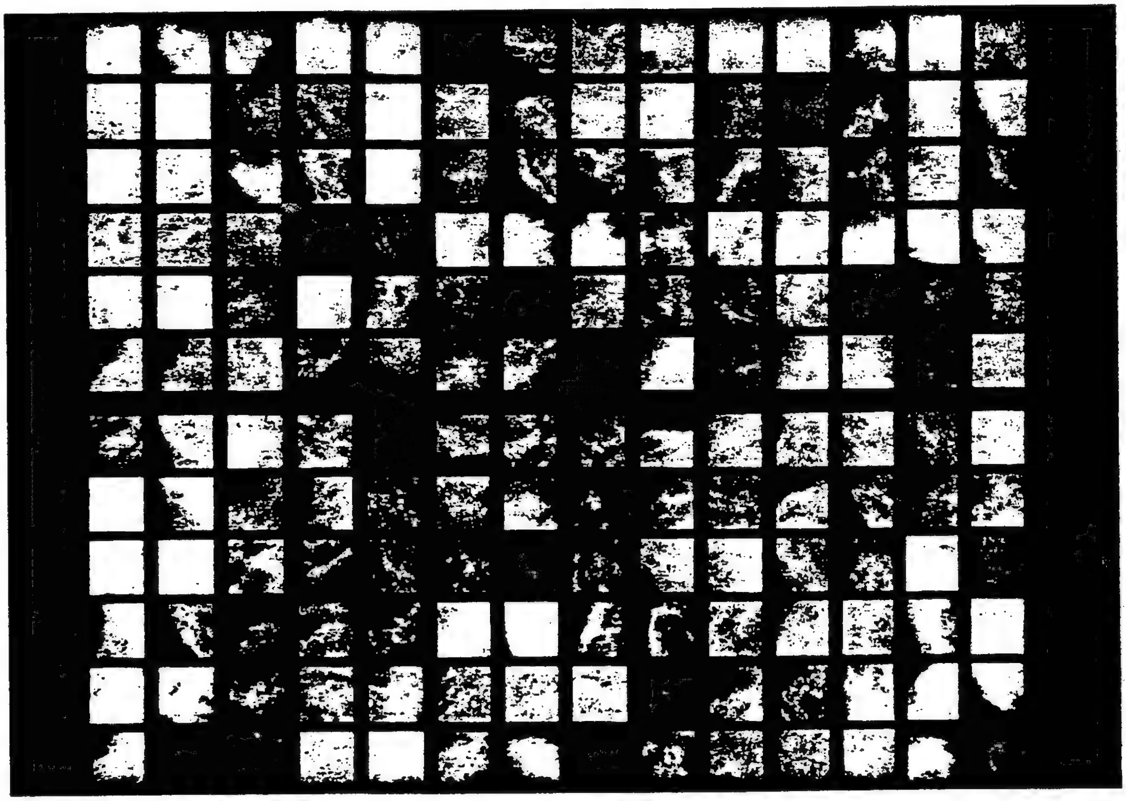
B. Texture features

The input images were digitized to 12 bits of resolution. The average gray level of each ROI of the images was removed and replaced by a constant for all the ROIs before the texture analysis and wavelet transform were performed in order to reduce the variability of the texture features caused by exposure conditions. The texture features were calculated based on the SGLD matrix, also known as the concurrence or co-occurrence matrix.[23,24] The (i,j)-th element of the SGLD matrix, $p_{d,\theta}(i,j)$, is the joint probability that the gray levels i and j occur in a direction θ at a distance of d pixels apart (d is the distance in terms of number of pixels and is referred to as pixel distance in the following discussion) over the entire ROI. The joint probability describes the frequency that a pair of gray level values occurs between pixel pairs with a defined, relative spatial relationship. The SGLD matrix is a two-dimensional histogram. The matrix size depends on the gray level resolution (i.e., the bit depth) of the digitized image and the bin width used in determining the histogram. If the gray level resolution is n bits and the bin width is b gray levels, then the size of the SGLD matrix will be $a\times a$, where $a=2^n/b$. For example, for a 12-bit image, the matrix size of an SGLD matrix constructed with a bin width of 1 gray level is 4096×4096. The matrix size is reduced to 256×256 if a bin width of 16 gray levels is used. The increased bin width is equivalent to reducing the gray level resolution of the 12-bit image to 8 bits by eliminating the 4 least significant bits and using a bin width of 1 gray level in determining the SGLD matrix. Based on the findings of our previous study,[25] 8-bit gray level resolution provided the best classification accuracy when texture features calculated at a fixed pixel distance d were used. Therefore, 8-bit gray level resolution was chosen for the formulation of the SGLD matrices in this study.

Eight texture features were examined: correlation, energy, entropy, inertia, inverse difference moment, sum average, sum entropy, and difference entropy. Some of the texture features can be used to describe some visual properties of the images while others may be more abstract. For example, cor-



(a)



(b).

FIG. 1. The 168 case samples used in this study with ROIs containing (a) biopsy-proven masses, (b) dense breast tissue, (c) mixed dense/fatty breast tissue, and (d) fatty breast tissue. The upper halves are the G_1 cases and the lower halves the G_2 cases.

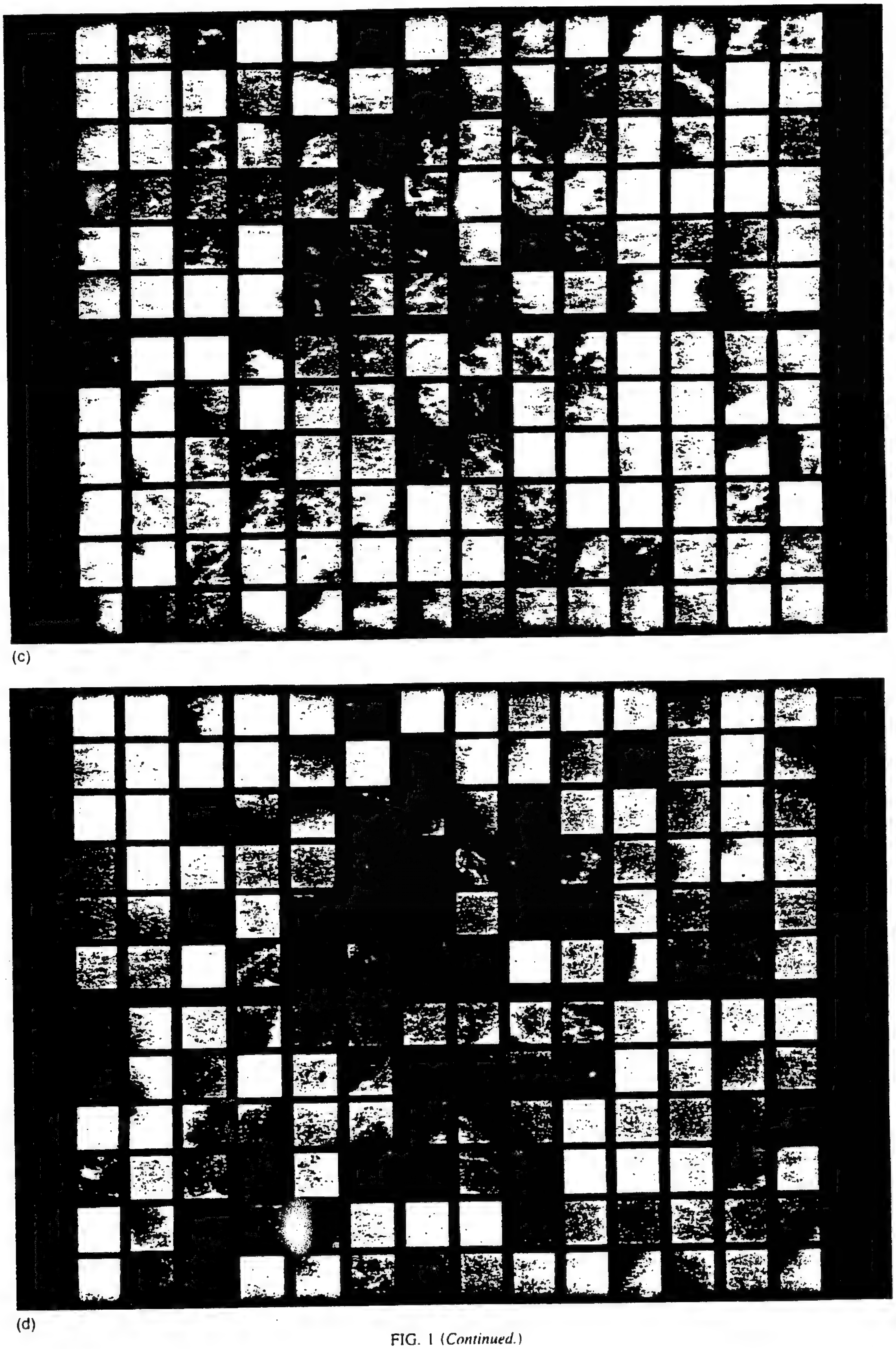


FIG. 1 (*Continued.*)

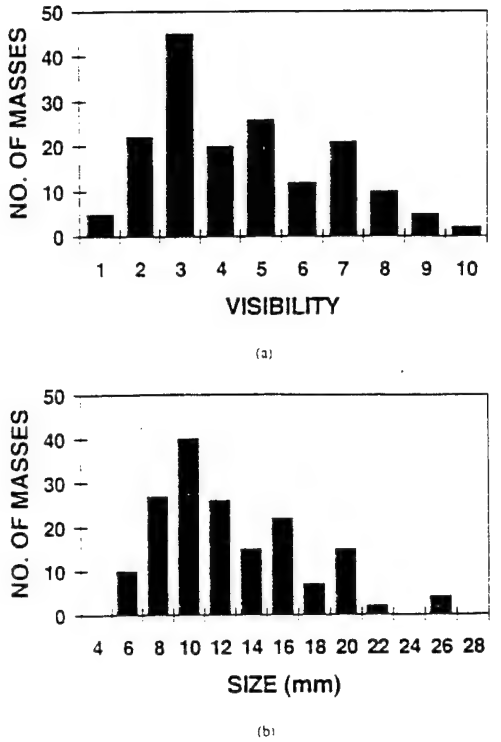


FIG. 2. The distribution of (a) the visibility score and (b) the size of the 168 masses.

relation is a measure of gray level dependency. Energy (or angular second moment) and entropy are measures of pixel homogeneity. Inertia (or contrast) represents the amount of intensity variation. It is difficult, however, to relate specific image characteristics to each of these features. The mathematical definitions of the features can be found in the literature[15,23,24] and are given in Appendix B.

Each texture feature was calculated at $\theta=0°$, 45°, 90°, and 135° for specified distances and/or scales. Since it is expected that the shape and the texture of masses in the ROIs do not have angular preferences, we averaged the features at $\theta=0°$, 90°, and at $\theta=45°$, 135°, and referred to these averaged features as features at 0° and 45°, respectively, in the following discussion. For a given pixel distance, the actual distance between the pixels on the image at 45° was equal to $\sqrt{2}$ times the actual distance at 0°. When the pixel distance increased, the differences in the actual distances between these angles become more significant. Because the texture features depended on the actual distance between the pixel pairs, the features at the two angles were treated separately in our multiresolution texture analysis.

C. Wavelet transform

The wavelet transform produces a multiscale representation of an image in which the geometric structures of the image are preserved within each sub-band or level. In Appendix A, we present a brief introduction to the wavelet transform. More details of the theory and applications can be found in the literature.[26–29]

Mallat presented a multiresolution framework with the discrete wavelet transform inherently embedded.[29] In this framework, the original image (Y) that has the highest resolution is referred to as level 0 ($i=0$) or scale 1 ($s=2^i|_{i=0}$). At scale 2, the original image is decomposed in the wavelet transform domain (similar to the spatial frequency domain in Fourier transform) into a low-pass sub-band image $Y_{2^i}^{\mathrm{LL}}|_{i=1}$ (referred to as approximation image at level 1 or scale 2, low-pass low-pass quadrant) and three bandpass sub-band images Y_2^{LH}, Y_2^{HL}, Y_2^{HH} (referred to as detail images in the low-pass high-pass, high-pass low-pass, and high-pass high-pass quadrants). At the next scale (scale 4), the approximation image at scale 2, Y_2^{LL}, is decomposed further into a low-pass sub-band approximation image Y_4^{LL} and three more bandpass sub-band images Y_4^{LH}, Y_4^{HL}, Y_4^{HH}. The decomposition can be stopped at some desired (lower) resolution or (larger) scale. Figures 3(a) and 3(b) illustrate the wavelet decomposition to level 4 or scale 16 of the ROIs containing a mass and normal parenchyma, respectively. The reconstruction of an image from the wavelet coefficients in the transform domain starts from the lowest resolution (largest scale) sub-band images.

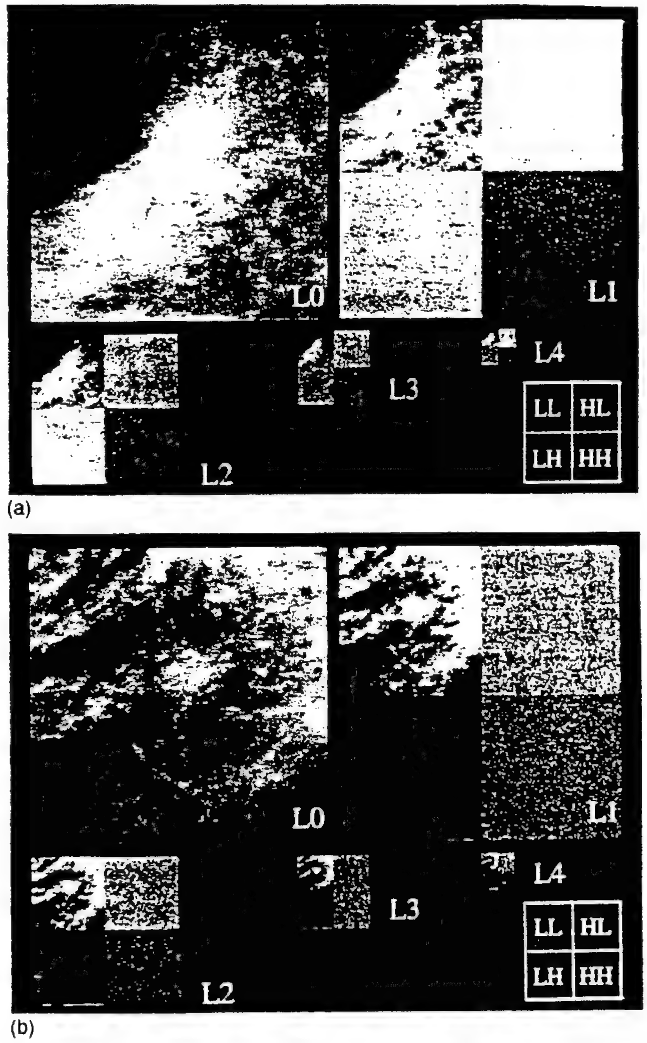


FIG. 3. Wavelet decomposition from level 0 ($L0$ or scale 1) to level 4 ($L4$ or scale 16) of (a) an ROI with a mass and (b) an ROI with normal breast tissue.

In this study, Daubechies' filter with four coefficients[27] was used as the wavelet filter for image decomposition (see Appendix A). A filter with a small number of wavelet coefficients was chosen because the width of the uncertainty band at the image boundary caused by convolution would be narrower. This allowed the decomposition to be performed to larger scales while still providing a sufficient number of usable pixels in the approximation image for the construction of an SGLD matrix. The chosen filter was also separable so that the fast wavelet transform algorithm could be employed in two-dimensional image analysis.

D. Multiresolution texture feature space

We used the original image Y (scale 1) and the low-pass sub-band approximation image $Y_{2^i}^{\mathrm{LL}}$ (scale 2^i, $i=1,\ldots,4$) to formulate SGLD matrices at multiple scales. The distance of the pixel pairs used at each scale was one pixel. The decomposition stopped at scale 16 so that the approximation image in the transform domain had 16×16 pixels. Effectively, the pixel distances of SGLD matrices formulated in this way at scales of 1, 2, 4, 8, and 16 corresponded to pixel distances of 1, 2, 4, 8, and 16 in the original image. A total of 80 features were calculated from each ROI (8 features×2 angles×5 levels) in this feature space. These 80-dimensional feature vectors based on the wavelet transform were denoted as $\mathbf{F}_{\mathrm{WT}}$.

As the scale in the wavelet transform increased, the statistical fluctuations in the SGLD matrices based on the smaller and smaller images could not be neglected due to the random sample errors. To reduce the statistical error in the SGLD matrices, we decomposed the original ROIs by wavelet transform to scale 4 so that the smallest image size was 64×64 pixels. Then the wavelet filter was applied once more without downward sampling. The resulting wavelet coefficients were obtained at scale 8 and were overcomplete and redundant.[29,30] However, this allowed the number of pixels used to construct the SGLD matrices to be kept at 64×64. The SGLD matrices at scale 8 were then constructed with distances of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12. These distances between pixel pairs were equivalent to the distances of 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, and 48 in the original image. Therefore, a total of 224 features were calculated from each ROI [8 features×2 angles×(1 pixel distance at scales 1, 2, 4+11 pixel distances at scale 8)] in this feature space. The feature vectors in this 224-dimensional feature space were based on wavelet transform and variable distances, and were denoted as $\mathbf{F}_{\mathrm{WV}}$.

To evaluate the effect of the wavelet transform on the classification results, we compared the features described above to those extracted from the SGLD matrices of the original image. The SGLD matrices were constructed with pixel distances of 1, 2, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, and 48. These distances corresponded to those used in the calculation of $\mathbf{F}_{\mathrm{WV}}$ when the latter were converted to equivalent pixel distances in the original image. Therefore, a total of 224 features were calculated from each ROI (8 features×2 angles×14 pixel distances) in this feature space. These feature vectors based on SGLD matrices from the original images with variable distances were denoted as $\mathbf{F}_{\mathrm{VD}}$. From the 224 features, we could also select a subset of 80 features at $d=1$, 2, 4, 8, and 16. The pixel distances in this subset corresponded to the pixel distances used for the calculation of the features in $\mathbf{F}_{\mathrm{WT}}$. The 80-dimensional feature vectors obtained from this subset of features with variable distances were denoted as $\mathbf{F}_{\mathrm{VDS}}$. The impact of the wavelet transform on the discriminant power of the texture features was studied by comparing the classification results obtained with $\mathbf{F}_{\mathrm{VD}}$ and $\mathbf{F}_{\mathrm{VDS}}$ to those obtained with $\mathbf{F}_{\mathrm{WV}}$ and $\mathbf{F}_{\mathrm{WT}}$, respectively.

E. Linear discriminant analysis

Linear discriminant analysis[31] is a systematic statistical technique to classify individuals or cases into one of the mutually exclusive classes based on certain indices or predictor variables. These indices or predictor variables may have certain correlations with one another. In a two-class classification problem, for example, a linear combination of these variables is formed and the coefficients are determined based on certain optimization criteria. One of such criteria, proposed by Fisher, is that the ratio of the difference of the means of the linear combination in the two classes to its variance is maximized.[31,32]

The discriminant analysis in the SPSS software package [M. J. Norusis, *SPSS for Windows Professional Statistics*, Release 6.0 (SPSS Inc., Chicago, IL, 1993)] was used in this study. The extended feature spaces as explained above were each used as a pool of predictor variable candidates for a two-class discriminant analysis that contained a mass class and a normal tissue class. Similar to the situation of multiple linear regression, including a large number of possible predictor variables in the linear model of the discriminant function is not a good strategy. Inclusion of irrelevant variables will not improve the classification accuracy and will decrease the generalization capability of the classifier. Because of the large number of features in the pools, it is a formidable task to test all different feature combinations at different numbers of feature variables to find the best combination. Therefore, we utilized a stepwise feature selection procedure to select predictor variables in each feature space. Five selection criteria are provided in the SPSS package, including (1) the minimization of Wilks' lambda, (2) the minimization of unexplained variance, (3) the maximization of the between-class F statistic, (4) the maximization of Mahalanobis distance, and (5) the maximization of Lawley–Hotelling trace (Rao's V). For each feature space, we tested all available selection criteria. With each criterion, we performed stepwise feature selection on all the 168 cases using the program default values for the inclusion and exclusion threshold parameters and the termination criterion. The selection criterion that provided the best classification result would be chosen. Since the program default values of the parameters might not be the optimal choices for our application, we varied the parameter values of the chosen criterion in an attempt to further improve the classification results. For our data sets, when the thresholds were set higher than the default values, fewer feature variables would be included and the classification accuracy decreased. When the thresholds were set lower than the default values, more features would be included and the classification results might improve. However, when the thresholds were lowered further and too many features were included, the classification would deteriorate. The set of feature variables that provided the best classification in this selection process were used for the formulation of the discriminant function in the given feature space. For simplicity, we will refer to this stepwise selection procedure with different thresholds as a stepwise or automatic selection process. Our feature selection process was by no means exhaustive. However, it would represent the best selection achievable within reasonable computational requirements.

To evaluate the capability of generalization of a trained classifier, we randomly divided the 168 cases into two groups (G_1 and G_2) of equal size. We used the features selected with the procedure described above as discriminant variables. If a given group was used for training, the feature values of each case from that group were used to optimize the coefficients of the linear discriminant function. The training cases were then classified with the linear discriminant function as a verification of consistency. The other group was used as test cases of which the feature values were input to the classifier and the discriminant score of each case was calculated from the linear discriminant function. One of the two groups was alternately used as the training group so that the variability of the classifier with different training groups could be observed.

Receiver operating characteristic (ROC) analysis[33,34] was used to evaluate the overall performance of the linear discriminant functions, in addition to the classification results reported by the SPSS program under certain prior probability assumptions. For a two-class problem, the ROC curve could be obtained using the Bayes' rule by changing the prior probability. Alternatively, the discriminant score from the canonical discriminant function could be used as the decision variable in the ROC analysis. Figure 4 demonstrates such a distribution of discriminant scores based on the linear combination of features calculated from wavelet coefficients at variable distances. The distribution of the discriminant scores of the ROIs in the training or the test group was input into

TABLE I. Texture features selected by stepwise discriminant analysis.

(a) From $\mathbf{F}_{WT}$ and $\mathbf{F}_{VDS}$

scale	1		2		4		8		16	
pixel distance	1		2		4		8		16	
θ	0°	45°	0°	45°	0°	45°	0°	45°	0°	45°
correlation							Δ		•	•
difference entropy									•	•
energy		•								
entropy										
inertia		•			•	•			•	
inv. dif. moment								•		
sum average	Δ			Δ			•	Δ	•	Δ
sum entropy										•

(b) From $\mathbf{F}_{VD}$ and $\mathbf{F}_{WV}$

scale	1	2	4		8																	
Distance	1	2	4		8		12	16	20		24		28	32		36		40	44		48	
θ	0°	45°	0°	45°	0°	45°	45°	45°	0°	45°	0°	45°	0°	0°	45°	0°	45°	0°	0°	45°	0°	45°
correlation					□▲		□▲				□▲		□	▲		■□		▲	□	■□	▲	▲
dif. entropy	■		■▲			■						□			□		■□		□		■▲	□
energy						□										□						□
entropy	▲	▲			□		□	□						■	□						□	■□▲
inertia																■		▲				
inv. dif. moment						■□	▲	■		▲								□			■□	
sum average				▲		■	▲		□						▲		■		□	■▲	■	
sum entropy	□												■					□		■		□▲

(a) ● 13 features (automatic)selected from $\mathbf{F}_{WT}$. ∇ 5 features (automatic) selected from $\mathbf{F}_{VDS}$. Note: 0° represents the average of features at 0° and 90°; 45° represents the average at 45° and 135°.
(b) ■ 19 features (automatic) from $\mathbf{F}_{WV}$. □ 29 features (semiautomatic) from $\mathbf{F}_{WV}$. ▲ 20 features (automatic) from $\mathbf{F}_{VD}$. Note: Some distances/angles are not shown if no feature was selected. 0° represents the average of features at 0° and 90°; 45° represents the average at 45° and 135°.

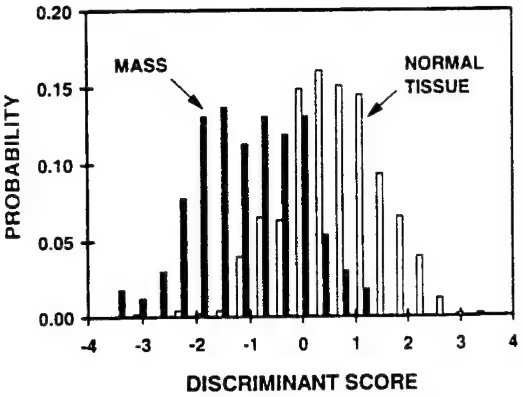


FIG. 4. An example of the probability density distribution of the discriminant scores of the masses and normal tissue. The discriminant scores were calculated from the canonical discriminant function that was optimized with all 672 ROIs with 19 features selected from multiresolution texture feature space $\mathbf{F}_{WV}$.

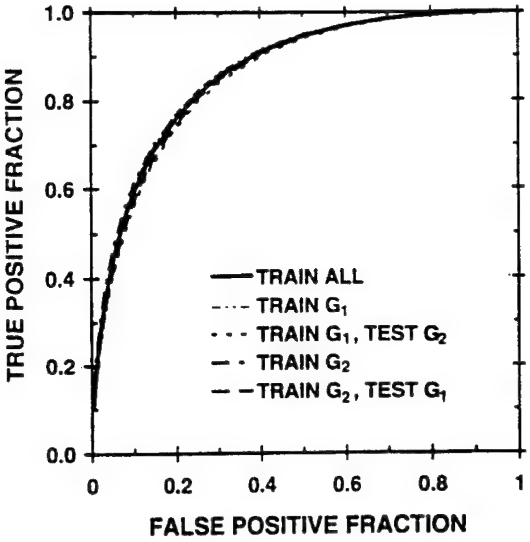


FIG. 5. ROC curves for classifying masses from normal tissue with discriminant function based on 13 features selected from texture feature space $\mathbf{F}_{WT}$.

the LABROC1 program.[35] which provided a maximum-likelihood estimation of a binormal ROC curve for training or testing, respectively. The area under the fitted ROC curve, A_z, was used as a performance index for evaluating the different sets of features selected from different multiresolution feature pools. The standard deviation (SD) of A_z estimated by LABROC1 was also reported. The CLABROC program was employed to test the statistical significance of the difference between A_z values of different sets of selected features.[36] The two-tailed p values were reported in the following comparisons.

III. RESULTS

A. Texture features based on wavelet coefficients

Stepwise feature selection was performed with the multiresolution texture features extracted from the feature space $\mathbf{F}_{WT}$. Thirteen features were selected as shown in Table I(a). The A_z and the estimated SD of the ROC curves are summarized in Table II. Figure 5 shows the ROC curves for the classification using the features derived from the wavelet coefficients. The A_z values of 0.858 and 0.854 for testing of G_1 and G_2, respectively, are higher than those of 0.817±0.027 ($p=0.02$) and 0.829±0.026 ($p=0.10$) obtained with texture features calculated from the SGLD matrix at a single distance of 20 pixels.[25]

B. Texture features based on original images with variable distances

To evaluate whether the improvement of classification over the results using features based on a single distance[25] is caused by the low-pass filtering in the wavelet transform or by the changes in the pixel distances, we used the same 13 features variables selected from $\mathbf{F}_{WT}$ but the feature values were calculated from the SGLD matrices based on the origi-

TABLE II. Comparison of the area under the ROC curves, A_z, obtained from different feature spaces.

Number of Features	Feature Space	Features extracted from scales	Training on G_1 and G_2	Training on G_1 Testing on G_2		Training on G_2 Testing on G_1	
			A_Z (Train)	A_Z (Train)	A_Z (Test)	A_Z (Train)	A_Z (Test)
13*	$\mathbf{F}_{WT}$	1, 2, 4, 8, 16	0.864±0.016	0.869±0.021	0.854±0.023	0.868±0.022	0.858±0.022
13◆	$\mathbf{F}_{VDS}$	1	0.796±0.019	0.808±0.026	0.781±0.027	0.798±0.027	0.787±0.027
5*	$\mathbf{F}_{VDS}$	1	0.758±0.021	0.766±0.028	0.747±0.029	0.754±0.029	0.760±0.028
20*	$\mathbf{F}_{VD}$	1	0.885±0.014	0.834±0.024	0.837±0.024	0.905±0.018	0.857±0.022
19▼	$\mathbf{F}_{VD}$	1	0.871±0.015	0.883±0.019	0.836±0.025	0.878±0.021	0.859±0.022
19*	$\mathbf{F}_{WV}$	1, 2, 4, 8	0.884±0.014	0.899±0.018	0.853±0.025	0.887±0.021	0.859±0.022
29△	$\mathbf{F}_{WV}$	1, 2, 4, 8	0.887±0.014	0.904±0.018	0.840±0.026	0.903±0.018	0.855±0.022

*Automatic feature selection.
◆ Features corresponding to those automatically selected from $\mathbf{F}_{WT}$.
△Features corresponding to those automatically selected from $\mathbf{F}_{WV}$.
▼Semiautomatic feature selection.

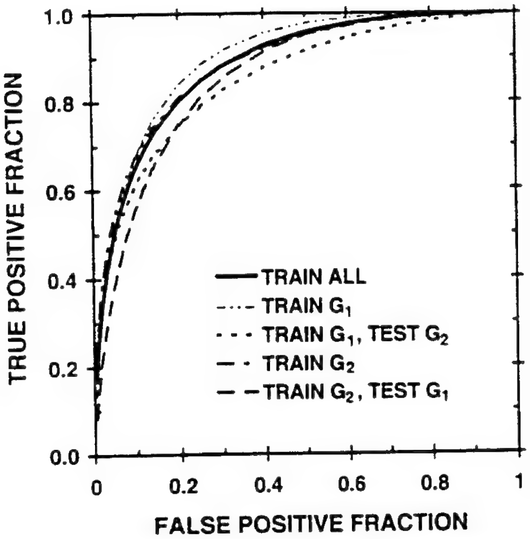


FIG. 6. ROC curves for classifying masses from normal tissue with discriminant function based on 19 features selected from texture feature space $\mathbf{F}_{WV}$.

nal images at equivalent distances, $\mathbf{F}_{VDS}$. The A_z values of the ROC curves with the same training and test groups (Table II) are significantly lower than the corresponding A_z values ($p<0.006$) with features extracted from the wavelet coefficients.

Stepwise feature selection was also performed on the entire data set of 168 cases from the feature space $\mathbf{F}_{VDS}$. The five features selected are listed in Table I(a). When this set of features was used to formulate the discriminant function, there was no improvement in A_z (Table II), compared with the results using the 13 features with feature values from the same $\mathbf{F}_{VDS}$ space. The differences between the A_z values obtained with 13 features from $\mathbf{F}_{WT}$ and the corresponding A_z values obtained with 5 features from $\mathbf{F}_{VDS}$ were statistically significant ($p<0.0002$). When the entire feature space of $\mathbf{F}_{VD}$ was used in the stepwise feature selection, 20 features were selected as listed in Table I(b). The A_z values for classification in both the training and test groups are significantly higher than those obtained with 5 features from $\mathbf{F}_{VDS}$ ($p<0.025$). As can be seen from Table I(b), 12 out of the 20 features were selected from distances greater than 16 pixels. This indicates that the information at larger distances which is not present in $\mathbf{F}_{VDS}$ is important in the classification of mass and normal tissue. Some of the A_z values obtained with these 20 features from $\mathbf{F}_{VD}$ are higher than those with 13 features from $\mathbf{F}_{WT}$, while the others are lower than those obtained from $\mathbf{F}_{WT}$, with p values ranging from 0.06 to 0.77. This is an indication that the discriminant power of the features from $\mathbf{F}_{VD}$ is comparable to that of the features from $\mathbf{F}_{WT}$.

C. Texture features based on wavelet coefficients at variable distances

Figure 6 illustrates the ROC curves for training and testing when stepwise feature selection was performed on the texture features extracted from the feature space $\mathbf{F}_{WV}$. The 19 features selected are listed in Table I(b). As shown by the A_z values in Table II, when the selected features were used to formulate the discriminant function, the classification results for the training sets improved in general, with p values ranging from 0.08 to 0.40, whereas the test results were almost the same as those obtained with the 13 features from $\mathbf{F}_{WT}$, with p values of 0.93 and 0.88. As can be seen from the same table, if these 19 variables were used on the feature values from $\mathbf{F}_{VD}$, the A_z values were similar to or slightly lower than those obtained with $\mathbf{F}_{WV}$. The differences are statistically significant for A_z (training on G_1 and G_2) and for A_z (testing on G_2) at $p<0.03$, and are insignificant for the other A_z values with p values ranging from 0.23 to 0.60.

We also selected the features in two steps, referred to as semiautomatic selection. First, we input texture features of the same type, e.g., correlation, calculated at all scales and distances into the discriminant analysis program. By using the stepwise selection method with reduced thresholds for the F values for variable entry and removal, we found the scales and distances that are important for classification for each texture feature. Then we applied the stepwise procedure again to all features at their selected scales and distances to further reduce the number of features. In this way, 29 features were selected as shown in Table I(b). Although most of them were different from the 19 features selected automatically, the overall classification results did not show much difference, indicating that some of the features used in one discriminant function might be linearly correlated with some of the features in the other discriminant function. The classification results (Table II) improved slightly in the training groups (with p values ranging from 0.08 to 0.74) but deteriorated in the testing groups (with p values of 0.24 and 0.54), probably because the increased number of features used in the discriminant function limited its capability for generalization.

IV. DISCUSSION

A. Multiresolution texture analysis

Textures are generally recognized as being fundamental to perception, although there is no precise definition or characterization of textures available in practice. Intuitively, texture descriptors provide measures of properties such as smoothness, coarseness, and regularity. When an image is composed of elements of texture primitives, the description of the image by texture features can be very effective. One of the advantages is that the texture features are shift invariant and can be made orientation invariant by averaging over various angles. This is very important since the location and orientation of the mass in the ROI can be arbitrary.

The masses found in clinical mammograms have very different shapes and sizes. It is a challenge to find a universal feature or a set of features that can differentiate the masses from the normal tissue and parenchymal structures in the breast. It is also difficult to define *a priori* an optimal resolution for the ROIs. A multiresolution approach could provide a scale-invariant interpretation of an image.

The wavelet transform is closely related to the well-known Fourier transform through the short-time Fourier transform or Gabor transform. It is considered a natural way of decomposing the image energy into different frequency bands through convolution with the translated and dilated version of a function called the "mother wavelet."[29] Unlike the Fourier transform where the coefficients in the transform domain do not reflect the local spatial variations, the wavelet coefficients retain the spatial variations of the original image.

In the multiresolution framework using wavelet decomposition proposed by Mallat,[29] the transform domain contains a minimum set of coefficients from which the reconstruction of the decomposed image is perfect or lossless. In the successive image decomposition, the approximation image in the current scale is decomposed into an approximation image and three detail images in the larger scale. Once the mother wavelet is chosen, the coefficients, which contain one approximation image and a series of detail images at different scales, are nonredundant and the transform is one-to-one. The extraction and condensation of image information through Mallat's framework are very efficient. Thus the wavelet transform is often used for image compression.[18,37] In the classification and pattern recognition problem, however, the focus is on the extraction of those features that can provide maximum distinction among different classes rather than on the minimal representation of the original image. In our current texture analysis, we used the approximation images at different scales, which are redundant representations of the original image. Such representations may be helpful in classification and pattern recognition applications, as demonstrated by the improvement in classification accuracy in comparison to the results obtained with features at a single distance,[25] or to the results obtained with features at variable distances without wavelet transform.

The discrete wavelet transform can be described as a cascaded process with two basic operations: filtering and down sampling. There are certain requirements for a filter to be wavelet filter.[26] Although it is possible to find optimal wavelet filters for certain types of images, our focus in this work is on the feasibility of multiresolution features for classification of masses from normal tissue rather than the optimization of this procedure. Therefore, an orthonormal four-weight Daubechies' filter with compact support[27] was used for our image decomposition. When the down-sampling process effectively reduces the image size by a factor of 2 in each direction as the scale increases, the reduced size of the distortion at the boundary will help keep as much useful image information as possible for texture calculation.

B. Comparison of classification accuracy with features from different feature spaces

To compare the discriminant power of texture features calculated from the wavelet coefficients to those from the original images, we used the feature variables with equivalent distances ($\mathbf{F}_{VDS}$). Using the features selected by the stepwise procedure, the classification results based on the features from $\mathbf{F}_{WT}$ were significantly better than those based on the features from $\mathbf{F}_{VDS}$. If we used the 13 features automatically selected from $\mathbf{F}_{WT}$ but formulated the discriminant functions based on the texture feature values from $\mathbf{F}_{VDS}$, the classification results demonstrated similar differences. This indicates that the texture features at equivalent distances from the wavelet transform domain have better discriminant power than those from the original images. However, when texture features up to distances of 48 ($\mathbf{F}_{VD}$, corresponding to 4.8 mm for 0° features and 6.79 mm for 45° features) are available for feature selection, the discriminant power of the texture features from the original images can reach as high as that of the features from $\mathbf{F}_{WT}$ or $\mathbf{F}_{WV}$. As can be seen from the features selected from each space shown in Tables I(a) and I(b), the texture information at large distances is important for the classification task. The feature space $\mathbf{F}_{VDS}$ does not provide such important information, resulting in poor classification. On the other hand, although the features in $\mathbf{F}_{WT}$ were calculated at distances equivalent to those of $\mathbf{F}_{VDS}$, the low-pass filtering effectively increases the correlation distances of the features. The structural information and energy of the original image obtainable at larger distances than the maximum equivalent distance of 16 pixels are condensed into the wavelet coefficients used for the calculation of $\mathbf{F}_{WT}$. The fact that the features from $\mathbf{F}_{WV}$ do not provide significant improvement (at least for the test groups) in the classification results indicates that the compression of image information is efficiently accomplished by the wavelet transform so that the additional information in $\mathbf{F}_{WV}$ is redundant as expected.

The overall operations of the discrete wavelet transform can be summarized as bandpass filtering (including low-pass and high-pass filtering) and downward sampling (decimation). The approximation images with the wavelet coefficients are the result of the low-pass filtering from convolution with the orthogonal scaling function.[29] The detail images obtained through convolution with the orthogonal wavelet function contain the edge (or high-frequency) information of the images. The texture features based on the multiresolution approximation images demonstrate improvement compared with those based on the original images for the classification of masses from normal tissue. This seems logical since, unlike microcalcifications that contain high-frequency components, the masses usually have relatively lower frequency contents. The frequency components of the background normal tissue are also in the low-frequency region, which makes the differentiation much more difficult. As the scale increases (by downward sampling or by increasing the distances in SGLD matrix formulation), the spatial resolution becomes lower while the low-frequency bands becomes narrower. The texture features based on the wavelet coefficients with decreasing low-frequency bandwidth demonstrate statistical difference between masses and normal tissue. At the same time, the effect of the noise with relatively high frequency is eliminated. The subtle differences between the masses and the normal tissue in the low-frequency range are therefore revealed when the difference in the changes of the low-frequency bands between them is utilized through multiresolution analysis. This may explain our finding that classification results with the multiresolution textures are better than those with single distance textures,[25] except for the results obtained with features selected from $\mathbf{F}_{VDS}$. It may be noted that the maximum distance of 16 pixels used in $\mathbf{F}_{VDS}$ is lower than the selected distances of 20 pixels in the single resolution texture analysis.

It is expected that the detail images in the wavelet transform domain contain valuable information about the difference between masses and normal tissue. When radiologists observe some large, suspicious structure, they will usually inspect it in more detail to determine whether it is a mass. However, we found that using the texture features based on the detail images in the wavelet transform domain to formulate the discriminant function did not result in proper classification. It seems that the statistical summary of the textures used here is not effective for the detail images. We will explore the use of other statistical features to extract the information contained in the detail images in future studies.

The reason that the features from the wavelet transform improve the classification results can also be explained as the result of the low-pass filtering operation. In this sense, other low-pass filters can also be used. This provides the possibility of designing optimal filters for the masses so that the classification results can be further improved. An advantage of the wavelet transform over other low-pass filters is that it provides an integral multiresolution framework with great computational efficiency.

The down-sampling process in the wavelet transform effectively reduces the number of pixels in the approximation image at each scale. The reduced size of the approximation images at larger scales will cause more variability in SGLD matrix formulation, thereby affecting the accuracy of the textures estimated at lower image resolution. As the scale increases, the statistical fluctuations of the SGLD matrix based on the smaller and smaller images cannot be neglected due to the random sample errors. In fact, when the approximation images at scale 32 with 8×8 pixels (equivalent to a pixel distance of 32) were used, the texture features did not show any differences between ROIs containing mass and ROIs containing normal tissue due to the small number of pixel pairs for the SGLD matrix formulation. To improve the statistical accuracy of the SGLD matrices, we used the information contained in the decimated coefficients in the wavelet transform and increased the number of discrete distances at which the SGLD matrices, thereby texture features in $\mathbf{F}_{\mathrm{WV}}$, could be calculated. Equivalently, this implies that features based on the information in the low-frequency bands with different bandwidths are used for classification. Although this did not significantly improve the classification results for the current data set, the features from $\mathbf{F}_{\mathrm{WV}}$ may be statistically superior to those from $\mathbf{F}_{\mathrm{WT}}$ because of the reduced uncertainties in the SGLD matrices.

C. Linear discriminant analysis

The classification accuracy is dependent on the feature variables in the linear discriminant function. We observed that when more features were used for the discriminant functions, there was a trend that the training results would improve at the expense of the test results. This is probably because the classifier has too many unknown parameters and is tuned toward the training group when it contains a small number of cases. The resulting discriminant function may not be representative for the general population. Therefore, the generalization capability of the classifier may deteriorate as the number of features used in the linear discriminant function increases. A similar situation arises when other classifiers, e.g., neural network, are used. We also observed that the feature variables selected by the stepwise discriminant analysis was dependent on the case samples in the training set. If we used the training subgroups to select feature variables, the feature variables selected from G_1 were not identical to those from G_2. Therefore, we used the whole data set (G_1 and G_2) to select the feature variables. As the number of case samples increased in the data set for feature selection, the statistical uncertainty of the distributions of the vectors in the feature space was reduced. This is expected to improve the robustness of the selected feature variables.

D. CAD application

One of our goals in the development of CAD methods in mammography is to assist radiologists in detection of suspicious masses on mammograms using computer vision techniques. Before the automated ROI detection method is fully developed, we used manually extracted ROIs to study the feasibility of using texture features for the classification of mass and normal tissue in different types of breast parenchyma. The results of this study demonstrated the potential of using multiresolution texture features for the classification task. The accuracy at an average A_z of 0.86 for the test sets represents a significant improvement over a single resolution approach.[25] Although further improvement in the accuracy is needed before clinical implementation, the algorithm can be incorporated into an automated mass detection program as a step to reduce false-positive ROIs. For example, we can set a decision threshold on the ROC curves (Fig. 6) at a true-positive fraction of 95% and a false-positive fraction (FPF) of 55%, thereby reducing 45% of the FPs while most of the true masses are retained. Alternatively, an accurate classification algorithm, once developed, can also be used independently from an automated detection algorithm. For example, it can be implemented in a CAD workstation and used by radiologists interactively to help differentiate ROIs indicated by the radiologists. The texture information used by the computer analysis may complement the human visual perception. The classification accuracy required, the best operating point on the ROC curve, and the appropriate approach of CAD implementation that can be most useful to radiologists are important topics of investigation in the future.

It is well known that the accuracy of a classifier for FP reduction depends on the specific types of FPs generated in the detection process, which may vary with different automated detection schemes or human observers. The accuracy may also depend to some extent on the properties of the image acquisition system used, such as the amplification mode, dynamic range, or spatial resolution. The coefficients in the linear discriminant function and the selected feature variables are expected to be different when the classifier is used in conjunction with different detection programs. The usefulness of this study lies in the fact that we developed a general approach to the extraction of multiresolution texture features and demonstrated their effectiveness in classification of masses and normal tissue. When this method is applied to a specific task, the classifier must be trained with ROIs representative of the population detected in that process, using the procedures developed in our study as a guide. It is also important that a much larger number of training samples than that used in this feasibility study is used in order to ensure the generalization capability of the trained classifier.

V. CONCLUSION

In this study, we examined the application of multiresolution texture features in the classification of masses and normal breast parenchyma. With linear discriminant analysis, we demonstrated that multiresolution texture features from the approximation images in the wavelet coefficients at different scales, $\mathbf{F}_{\mathrm{WT}}$, provide significant improvement in the classification accuracy over the features from the original images at equivalent distances, $\mathbf{F}_{\mathrm{VDS}}$. The features from the combination of wavelet coefficients and variable distances, $\mathbf{F}_{\mathrm{WV}}$, can further improve the classification accuracy, although the improvement falls short of statistical significance. The A_z under the ROC curve using 19 features from the $\mathbf{F}_{\mathrm{WV}}$ feature space reached an average of 0.89 for training and 0.86 for testing. The approach developed here can be incorporated into a CAD procedure which may assist radiologists in the detection of suspicious lesions on mammograms. While improvement in the classification accuracy is still necessary for clinical applications, our results demonstrate the feasibility of using multiresolution textures for the classification of masses from normal tissue on digital mammograms.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant No. CA 48129, a Faculty Research Award (FRA-334) from the American Cancer Society, and U.S. Army Grant No. DAMD 17-93-J-3007 (through Subgrant No. GU RX 4300-803UM from Georgetown University). The content of this publication does not necessarily reflect the position of the Georgetown University or the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for providing the LABROC1 and the CLABROC programs.

APPENDIX A: WAVELET TRANSFORM

In the following, we will briefly describe the basic approach of the wavelet transform that is related to this paper. For simplicity, one-dimensional wavelet transform is discussed. Generalization to two-dimensional space is straightforward.

In the wavelet transform, a signal $f(x)$ is decomposed with a family of real orthonormal bases $\psi_{j,n}(x)$ obtained through translation and dilation of a kernel function known as the mother wavelet:

$$\psi_{j,n}(x)=2^{-j/2}\psi(2^{-j}x-n), \tag{A1}$$

where j and n are integers. The wavelet coefficients of the signal $f(x)$ can be obtained through the decomposition

$$c_{j,n}=\int_{-\infty}^{+\infty}f(x)\psi_{j,n}(x)dx. \tag{A2}$$

The signal can be reconstructed from the wavelet coefficients $c_{j,n}$ and the wavelet bases $\psi_{j,n}(x)$ through the synthesis formula

$$f(x)=\sum_{j,n} c_{j,n}\psi_{j,n}(x). \tag{A3}$$

The mother wavelet $\psi(x)$ can be constructed from a scaling function $\phi(x)$, which satisfies the two-scale difference equation[26,27]

$$\phi(x)=\sqrt{2}\sum_k h(k)\phi(2x-k). \tag{A4}$$

The wavelet kernel $\psi(x)$ is related to the scaling function via

$$\psi(x)=\sqrt{2}\sum_k g(k)\phi(2x-k), \tag{A5}$$

where

$$g(k)=(-1)^k h(1-k). \tag{A6}$$

Several conditions have to be met in order for the set of wavelet functions in Eq. (A1) to be unique, orthonormal, and have a certain degree of regularity.[26] Different sets of coefficients satisfying those conditions can be found in the wavelet literature.[27–29]

In the discrete wavelet transform, fast recursive algorithms for wavelet decomposition have been developed. The pyramid wavelet algorithm, which we used for the multiresolution image analysis in this study, decomposes the signal into two parts in the next, larger scale: an approximation signal with the scaling function that has low-pass filter characteristics, and the detail signal with the wavelet function that has the bandpass filter characteristics. In our two-dimensional wavelet transform, we retained the coefficients that corresponded to the scaling function $\phi(x)$ at each scale for texture analysis.

APPENDIX B: TEXTURE FEATURES

An SGLD matrix element $p_{\theta,d}(i,j)$ is the joint probability of the gray level pairs i and j in a given direction θ separated by a distance of d pixels. For each ROI eight features were derived from its SGLD matrix of a given θ and d:

$$\text{energy}=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}p^2(i,j),$$

where n is the number of gray levels of the image;

$$\text{correlation}=\frac{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-\mu_x)(j-\mu_y)p(i,j)}{\sigma_x\sigma_y},$$

where

$$\mu_x=\sum_{i=0}^{n-1} i\sum_{j=0}^{n-1} p(i,j), \quad \sigma_x^2=\sum_{i=0}^{n-1}(i-\mu_x)^2\sum_{j=0}^{n-1}p(i,j),$$

$$\mu_y=\sum_{j=0}^{n-1} j\sum_{i=0}^{n-1} p(i,j), \quad \sigma_y^2=\sum_{j=0}^{n-1}(i-\mu_y)^2\sum_{i=0}^{n-1}p(i,j)$$

are the mean and variance of the marginal distributions $p_x(i)$ and $p_y(j)$, respectively;

morphological feature classifier and object splitting algorithm used in the segmentation method. The image database and the complete two-stage DWCE segmentation method is outlined in Section III. Finally, Sections IV and V contain the DWCE segmentation results along with a discussion of the advantages and limitations of the method.

II. Density-Weighted Contrast Enhancement Segmentation

We have developed a new algorithm using DWCE filtering with Laplacian–Gaussian (LG) edge detection for segmentation of low contrast objects in digital mammograms. The DWCE algorithm is used to enhance objects in the original image so that a simple edge detector can define the object boundaries. Once the object borders are known, morphological features are extracted from each object and used by a classification algorithm to differentiate mass and nonmass regions within the image.

A. *DWCE Preprocessing Filter*

Edge detection applied to the original digitized mammograms has not proven effective in detecting breast masses because of the low signal-to-noise ratio of the edges and the presence of complicated structured background. Fig. 1(a) shows a typical mammogram from our image database. It contains a single breast mass indicated by the arrow. This mammogram also contains dense fibroglandular tissue in the breast parenchyma. Although the mass is relatively obvious, the partially overlapping tissue makes the detection process difficult. In order to detect masses of varying shapes and intensities, we propose using an adaptive filtering technique to suppress the background structures and enhance any potential signals. Fig. 1(b) shows the preprocessed mammogram of Fig. 1(a). The background in this image is substantially reduced by the proposed adaptive filter, allowing object localization by a simple edge detector.

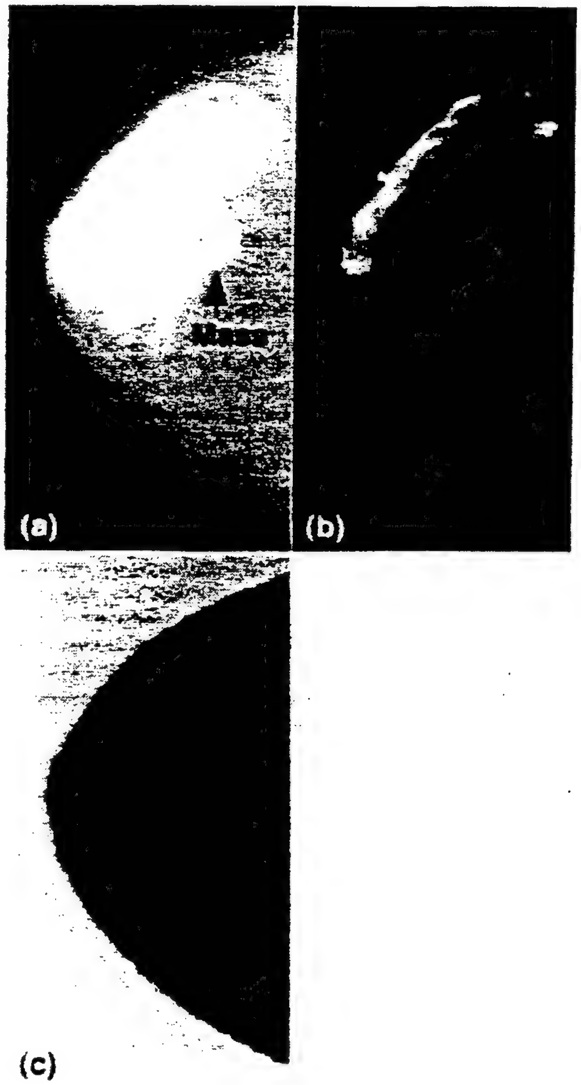


Fig. 1. (a) A typical mammogram from our image database, (b) the corresponding DWCE filtered image, and (c) the breast map defined by and used in the map rescaling.

The block diagram for the DWCE preprocessing filter is depicted in Fig. 2(a). It is an expansion of the local contrast and mean adaptive filter of Peli and Lim [19] designed for enhancing images degraded by cloud cover. The original image, $F(x,y)$, is initially passed through the map rescaler shown in Fig. 2(b). The rescaling first determines an estimate for the breast boundary i.e., the breast map, $F_{Map}(x,y)$. This is accomplished by rescaling $F(x,y)$ between 0.0 and 1.0 based on the maximum and minimum values within the whole image and then applying a single threshold. In our case, any image intensity value greater than or equal to 0.4 was considered part of the initial breast map estimate. All isolated objects were then identified using the Laplacian–Gaussian method described later in Section II-B. The region within the largest-area object was then selected as the final breast map, $F_{Map}(x,y)$. Fig. 1(c) depicts the detected breast map for the mammogram of Fig. 1(a). Using this breast map, the pixel values within the original image, $F(x,y)$, are again rescaled between 0.0 and 1.0. The rescaling range is now determined from the histogram of pixel values within the breast region. The pixel values defining the maximum and minimum of the rescaling range are set to be the maximum and minimum values containing at least 5% of the total pixel counts. This prevents outlying pixel values from skewing the rescaling.

The map rescaling produces a normalized image, $F_N(x,y)$, and allows a single set of filter parameters to be applied to all images in the set. $F_N(x,y)$ is next split into a density and a contrast image, $F_D(x,y)$ and $F_C(x,y)$, respectively. The density image is produced by filtering $F_N(x,y)$ with some type of low pass filter (e.g., local averaging, Gaussian smoothing, or median filtering). In the current DWCE filter implementation, zero-mean Gaussian smoothing with standard deviation, σ_D, is used. $F_D(x,y)$ thus directly correlates to a weighted average of the local optical density of the original film. The contrast image, $F_C(x,y)$, is also created by filtering $F_N(x,y)$, but the low pass filter is replaced with a bandpass or high-pass filter. In the current version of the DWCE, $F_C(x,y)$ is created by simply subtracting a Gaussian smoothed version

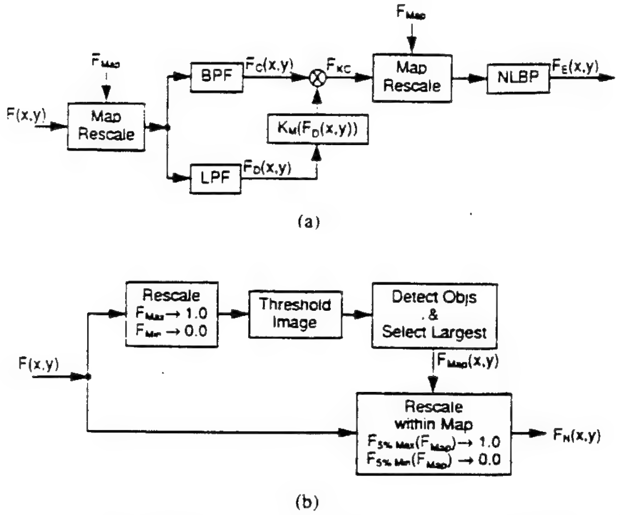


Fig. 2. (a) The block diagram of the DWCE preprocessing filter used for image enhancement and (b) the block diagram for the map rescaling.

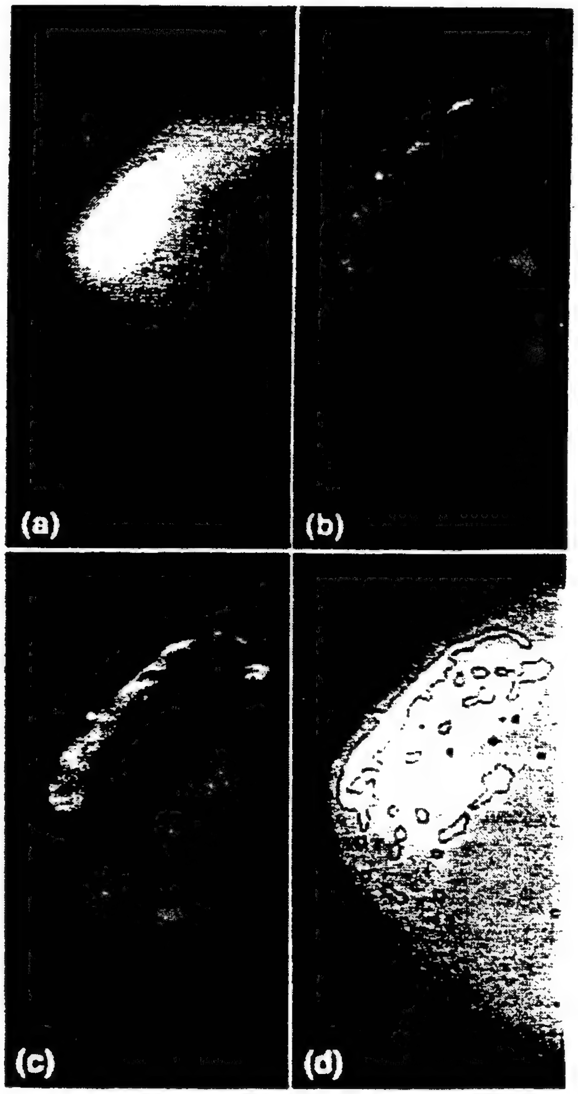


Fig. 3. (a) The density, $F_D(x,y)$, (b) contrast, $F_C(x,y)$, and (c) weight-contrast, $F_{KC}(x,y)$, images produced by the DWCE filter applied to the mammogram of Fig. 1(a) along with (d) the corresponding LG object image.

of $F_N(x,y)$ from itself

$$F_C(x,y) = F_N(x,y) - (G(x,y) * F_N(x,y)) \qquad (1)$$

where $G(x,y)$ is a Gaussian smoothing filter with zero mean and standard deviation, σ_C. This complement filter is selected to allow the entire frequency information of the normalized image to be used when $\sigma_D = \sigma_C$. Each pixel in the density image is then used to define a multiplication factor which modifies the corresponding pixel in the contrast image

$$F_{KC}(x,y) = K_M(F_D(x,y)) \times F_C(x,y). \qquad (2)$$

This is the essence of the DWCE algorithm. It allows the local density value of each pixel to weight its local contrast. Fig 3(a)–(c) show the density, contrast, and weighted contrast images, respectively, for the digitized mammogram of Fig. 1(a). The weighted contrast was created using the multiplication function shown in Fig. 4(a). Note that the DWCE filter substantially reduces the background and noise while retaining the significant breast structures.

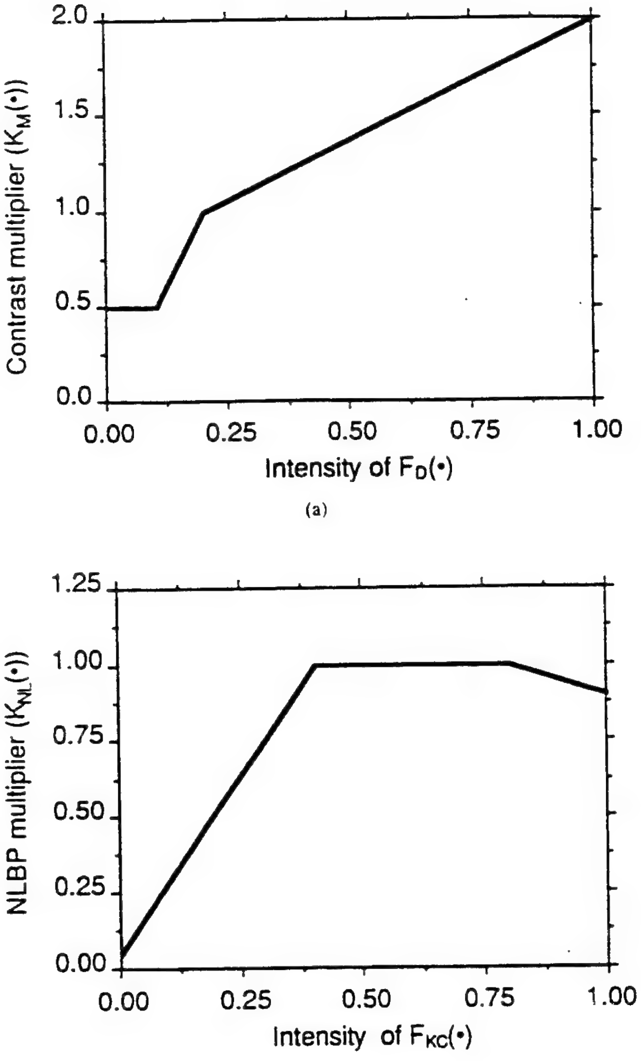
The output of the DWCE filter is a nonlinear rescaled version of the weighted contrast image. Each pixel in the weighted contrast image is used to define a second multiplication value, $K_{NL}(F_{KC}(x,y))$. The multiplication values are then multiplied by the weighted contrast of the corresponding pixels

$$F_E(x,y) = K_{NL}(F_{KC}(x,y)) \times F_{KC}(x,y). \qquad (3)$$

This produces the final filtered image, $F_E(x,y)$. Fig. 4(b) shows the nonlinear function, $K_{NL}(\cdot)$, used in the current DWCE implementation and Fig. 1(b) again shows the final enhanced image produced by this single DWCE filter stage.

The specific shapes for K_M and K_{NL} in Fig. 4 were determined experimentally by observing their affects on the detection. The shape of $K_M(\cdot)$ was selected to reinforce (i.e., $K_M(\cdot) \geq 1.0$) the contrast at pixels in $F_D(x,y)$ with medium to high intensity while reducing (i.e., $K_M(\cdot) < 1.0$) the contrast of low intensity pixels in the density image. The rationale is that only background is generally contained in the low intensity portion of $F_D(x,y)$ while masses and other breast structures will be seen at higher intensity values. Note $K_M(0.25) = 1.0$, so 75% of the intensity range will see contrast enhancement. This contrast multiplication function worked well in enhancing breast structures, but did not provide adequate separation between the structures. In order to isolate more of these structures and to help equalize the contrast across individual objects, a nonlinear rescaling was used. From Fig. 4(b) we can see that the very low contrasts are strongly deemphasized while the highest contrast range is slightly reduced. This rescaling sharpens the object borders by eliminating many of the low contrast edges that cause region merging and reduces the effect of extremely large contrasts on the edge detection.

The DWCE enhancement scheme is very general. It allows for great flexibility in defining the density and contrast im-



(a)

(b)

Fig. 4. Plots of (a) the weighted contrast multiplication function, $K_M(F_D(x,y))$, and (b) the nonlinear rescaling multiplication function, $K_{NL}(F_{KC}(x,y))$, used in the DWCE filter implementation.

ages by the selection of filter parameters. It also allows for significantly different degrees of enhancement based on the selected multiplication and nonlinear rescaling functions. In addition, we have found that the enhanced image is not very sensitive to small variations in $K_M(\cdot)$ and $K_{NL}(\cdot)$. Slight changes in the multiplication functions produced only slight visible variations in the image enhancement and had little affect on the detected edges. This was observed by Peli and Lim as well [19]. Since we defined the multiplication functions in the enhancement filter empirically, they may not be optimal. Therefore, variations in the multiplication functions of Fig. 4 (or completely new functions) may provide better overall performance.

B. Edge Detection

The primary motivation for enhancement prefiltering is to improve the detectability of mass edges. Edges are defined by changes or discontinuities in the intensity across an image. This feature is fundamentally important in image processing because it provides an indication of the physical extent of objects within the image. A common approach to monochrome edge detection is to apply linear or nonlinear edge enhancement followed by a threshold operation [20]. A simple example of edge enhancement is discrete differencing which is analogous to continuous spatial differentiation. However, with discrete differencing the edges are directionally dependent upon the differencing operation used. This dependence can be avoided by using a Laplacian mask which sharpens edges without regard to direction [20]. The performance of any edge detector can be severely degraded when an image is corrupted with noise. To alleviate this problem, statistical edge detection methods have been developed when the form of noise disturbance is known. Alternatively, edge detection algorithms have to be combined with smoothing filters to improve their performance.

In this study, object edges were detected from the DWCE prefiltered images using an "optimal" Laplacian–Gaussian (LG) edge detector. For a given image, $I(x,y)$, the LG edge detector defines edges as simply the zero crossing locations of

$$\nabla^2 G(x,y) * I(x,y) \tag{4}$$

where $G(x,y)$ is a two-dimensional Gaussian smoothing function [21]. The degree of smoothing is controlled by a single parameter, σ_E, the standard deviation of the smoothing function. This edge detector is optimal in the sense that the output energy near the edge features is maximized [22], [23]. It also tends to produce closed regions which makes detection of isolated objects within the image easier. The LG edge detector can be applied recursively from lower resolution (large σ_E) to high resolution (smaller σ_E) using the edge map of the previous stage as a guide for the current edge detector. The multi-resolution approach will improve the localization of the edges which are otherwise degraded by the Gaussian smoothing. In this study, we applied a single-stage LG edge detector with $\sigma_E = 2$ because the DWCE filter alone provided sufficient noise reduction.

Once the object edges have been defined in the image using the LG edge detection, each enclosed object is filled. This removes any holes that may have formed inside an object. Fig. 3(d) shows the enclosed filled edge regions produced by extracting the LG edges from the DWCE filtered image of Fig. 1(b). Each of the regions produced by the filling is defined by its edge pixels, thus forming a set of detected objects. This set of objects defines all the detected structures within the original breast image.

C. Object Splitting

One problem with the DWCE filter is that different structures within the breast can merge into a single connected region. The result is multiple objects merged into a single larger object. The morphological features of these large detected objects do not necessarily correlate to features of the smaller breast masses and normal tissue. In order to reduce the distortion due to merging, binary splitting is performed on the detected objects. The splitting algorithm searches

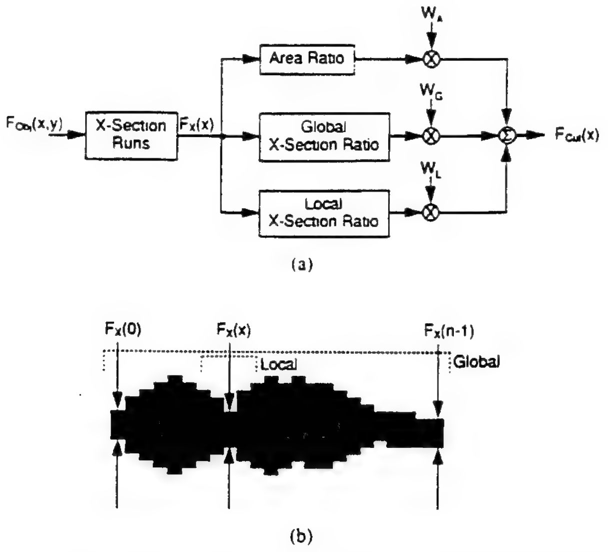


Fig. 5. (a) The block diagram for the splitting stage and (b) an example object containing the search ranges for the global and local cross-section ratios.

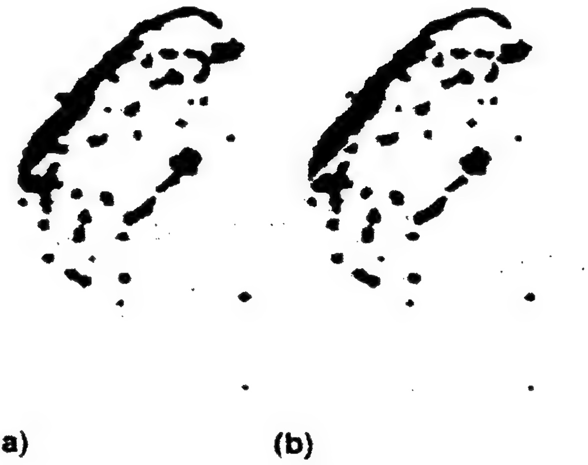


Fig. 6. (a) A set of detected objects and (b) the resulting set of objects produced by the splitting algorithm.

for narrowings in the cross section of objects (i.e., necks). Fig. 5(a) depicts the block diagram for the splitting algorithm. The algorithm initially finds the cross section width for each column in the object as shown in Fig. 5(b). This produces $F_X(x)$ which is a vector of length n. In the next stage of the splitting algorithm, the area ratio and the global and local cross-section width ratios are calculated for each column of the object. These values are defined as

$$F_{Area}(x) \equiv \frac{\min(A_R(x), A_L(x))}{\max(A_R(x), A_L(x))} \tag{5}$$

$$F_{Gbl}(x) \equiv \left\{ 1.0 - \frac{F_X(x)}{\max(F_X(z))} : z \in [0, n-1] \right\} \tag{6}$$

$$F_{Lcl}(x) \equiv \left\{ 1.0 - \frac{F_X(x)}{\max(F_X(z))} : z \in [x-2, x+2] \right\} \tag{7}$$

where $A_R(x)$ and $A_L(x)$ are the area of the right and left objects produced by splitting at location x, and the local and global ranges are shown in Fig. 5(b). At each potential neck location, x, a cut value is defined as a linear combination of the width of the cross-section ratios and area ratios of the two regions formed by the split

$$F_{Cut}(x) = W_G F_{Gbl}(x) + W_L F_{Lcl}(x) + W_A F_{Area}(x). \tag{8}$$

For the present study W_G, W_L, and W_A were chosen to be 1.5, 2.0, and 1.0, respectively. A maximum cut value for all narrowings in the vertical, horizontal, 45° and 135° directions is found and compared to a minimum cut threshold. If the maximum cut value exceeds this threshold the object is split at that point, otherwise, it is left unchanged. Fig. 6(a) and (b) shows a typical set of detected objects in the image before and after splitting, respectively. Round shaped objects were not cut while objects with necks were split in appropriate locations and directions. The advantage of this algorithm is that by incorporating the area ratio into the cut location, preference is given to neck locations near the center of the object. Note that this splitting algorithm is applied to the binary object images. The algorithm does not make use of information about the likelihood of multiple objects to define split locations.

D. Morphological Object Classification

Since the DWCE prefilter enhances both breast masses and normal tissue, a large number of detected objects are usually found. In order to reduce the number of objects to a manageable size, morphological features are extracted from each object and used in a preliminary screening of breast masses from normal tissue. The morphological features used in this classification include the number of edge pixels, area, shape and contrast of the objects. Two features, circularity and rectangularity, are used to characterize the shape of an object. To define these two features, the bounding box containing the object, and a circle with area equivalent to the object area and centered at its centroid location are first calculated. Fig. 7 shows the equivalent area circle, $F_{Eq}(x, y)$ with radius

$$R_{Eq} = \sqrt{\frac{\text{area}(F_{Obj})}{\pi}} \tag{9}$$

and the bounding box for the object, $F_{BB}(x, y)$. The definitions of circularity and rectangularity are then given by

$$\text{Circularity} \equiv \frac{\text{area}(F_{Obj} \cap F_{Eq})}{\text{area}(F_{Obj})} \tag{10}$$

$$\text{Rectangularity} \equiv \frac{\text{area}(F_{Obj})}{\text{area}(F_{BB})}. \tag{11}$$

Using these five morphological features classification was performed on the detected objects.

The morphological classification is not meant to be a final classification of the detected objects. Instead, it is used to reduce the number of objects so that further detailed analysis can be performed in each region. Once the number of objects have been reduced, regions of interest (ROI's) will be extracted based on the shape and location of the remaining objects. More

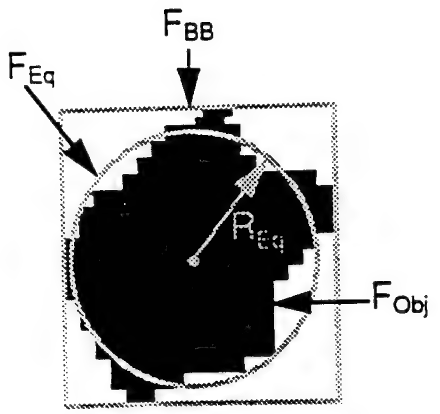


Fig. 7. An example object with its equivalent circle, F_{Eq}, and bounding box, F_{BB}, used in the definition of circularity and rectangularity, respectively.

sophisticated classification algorithms will then be applied to the extracted ROI's to differentiate between masses and normal breast tissue. Therefore, the goal in this classification block is to reduce the number of objects with minimal loss in the number of masses.

Simple thresholding, linear discriminant analysis (LDA), [24] and a back-propagation neural network (BPN) [25] have been investigated as potential classification schemes. Simple thresholding sets a maximum and a minimum value for each morphological feature. If a detected object falls within the bounds for each of the features, it is kept as a potential mass; otherwise it is considered to be normal tissue. LDA forms an optimal linear combination of the features which maximizes the group mean separation of the mass and nonmass objects. If the features follow a multivariate normal distribution with an identical covariance matrix for both groups, then LDA will yield the optimal classification. This linear combination forms a single discriminant score for each detected object [24], [26]. BPN also forms a single discriminant score for each detected object, but it finds the best nonlinear combination of features that minimizes the cost associated with misclassification. In our implementation, the BPN consists of an input layer, an output layer, and one hidden layer. Each layer contains a number of nodes interconnected to all nodes in the previous and the subsequent layers by weights. A weighted sum of node values from the previous layer stimulates a node in the subsequent layer through a nonlinear sigmoidal activation function. The neural network learns by supervised feedforward back-propagation training of the interconnecting weights [25]. A threshold applied to the LDA or BPN discriminant score provides a means for separating potential breast masses from the normal tissues. Discriminant scores above the threshold are classified as potential masses while scores falling below the threshold are considered as normal tissue and thus discarded.

III. Methods

A. Database

The clinical mammograms used in this study were randomly selected from the files of patients who had undergone biopsy in the Department of Radiology at the University of Michigan. The mammograms were acquired using a Kodak MinR/MRE screen/film system with extended cycle processing. The mammography systems have a 0.3-mm focal spot, a molybdenum anode, 0.03-mm-thick molybdenum filter and a 5:1 reciprocating grid. All systems have been certified by the American College of Radiology (ACR), and the image quality is monitored according to the ACR's recommended guidelines. Our selection criterion was simply that a biopsy-proven mass could be seen on the mammogram. Our data set in this preliminary study was composed of 25 mammograms. The size of the masses ranged from 6 mm to 26 mm with a mean size of 12.4 mm and included 11 malignant and 14 benign masses.

The mammograms were digitized with a LUMISYS DIS-1000 laser film scanner with a pixel size of 100 μm × 100 μm and 4096 gray levels. The DIS-1000 logarithmically amplifies the light transmitted through the mammographic film before digitization so that the gray levels are linearly proportional to optical densities in the range of 0.1 to 2.8 optical density units (O.D.). The O.D. range of the scanner was 0 to 3.5 with large pixel values in the digitized mammograms corresponding to low O.D. The digitized images are approximately 2000 × 2000 pixels in size. Before the DWCE segmentation was applied the images were smoothed within an 8×8 pixel window using local averaging and then subsampled by a factor of 8. This resulted in images of approximately 256×256 pixels for processing. Our data set was composed of 25 of the subsampled mammograms and will be referred to as the "subsampled" mammogram set in the following discussion.

B. DWCE Implementation

Fig. 8 shows the block diagram for the DWCE implementation used to detect breast masses in the 25 digitized mammograms. It was performed using two DWCE stages. In the first stage, the DWCE prefiltering, edge detection and simple thresholding classification were applied to each subsampled mammogram, and the potential mass objects were identified. For each potential mass object, an ROI was extracted from the corresponding subsampled mammogram using the bounding box of the object to define the region. The minimum size for the extracted ROI's was chosen to be 32×32 pixels. Any object with a bounding box smaller than this size had its bounding box uniformly expanded in each direction (horizontal and vertical) until it reached 32×32 pixels. This expanded bounding box was then used to define the extracted ROI region.

Each of the extracted object ROI's were then passed through a second DWCE stage. This stage included DWCE prefiltering, edge detection, object reduction, splitting and classification. The parameters used in the DWCE prefiltering, edge detection and object reduction steps were identical to those of their first stage counterparts. In the second stage classification, the simple thresholding, LDA or BPN (five input nodes, three hidden nodes and a single output node) classifiers were applied to the detected objects. This allows the three classification schemes to be compared. The detection accuracy was evaluated in terms of the number of true positives (TP's) for a given number of false

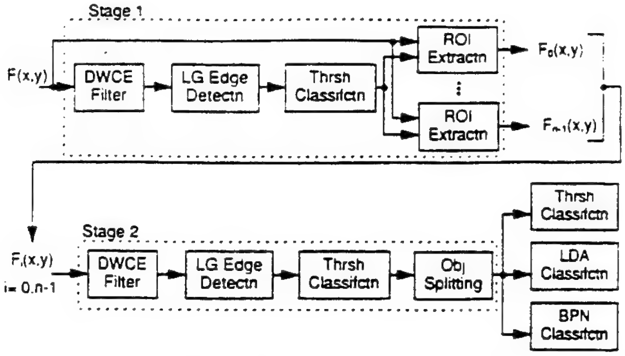


Fig. 8. The block diagram of the complete two stage DWCE segmentation method used for breast mass detection.

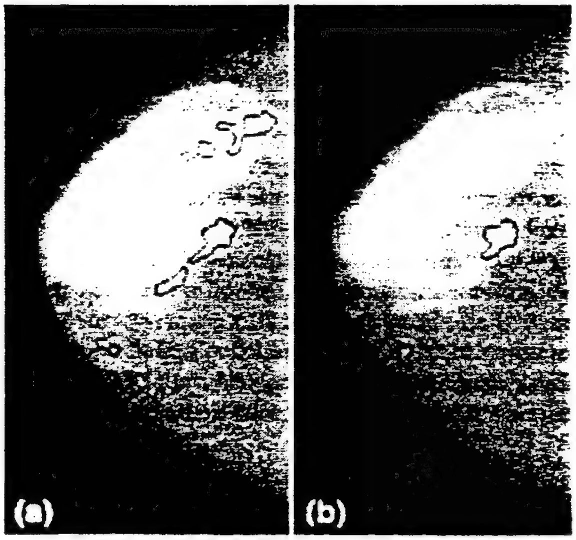


Fig. 9. (a) The detected objects obtained by the first stage and (b) second stage of the DWCE segmentation.

positive (FP) detections. A TP was considered as an object whose area overlaps the centroid of the biopsy proven mass as identified by a radiologist. Each mammogram in our data set contained a single TP object. All other objects classified as potential masses were considered as FP detections. Fig. 9(a) and (b) shows the detected objects from the mammogram of Fig. 1(a) after the first and second stages, respectively, in the DWCE segmentation. A simple thresholding classifier was used for object classification after both stages in this example.

IV. Results

The DWCE segmentation method was used to extract potential mass objects from the 25 subsampled mammograms. After the first stage of the DWCE segmentation, a total of 481 potential mass objects were detected including 24 of the 25 true mass objects. Local ROI's were then extracted using the object bounding boxes and passed through a second DWCE filtering stage. This second stage produced 218 detected objects which increased to 461 total objects after the splitting stage. The split object set again included 24 of the 25 true mass objects.

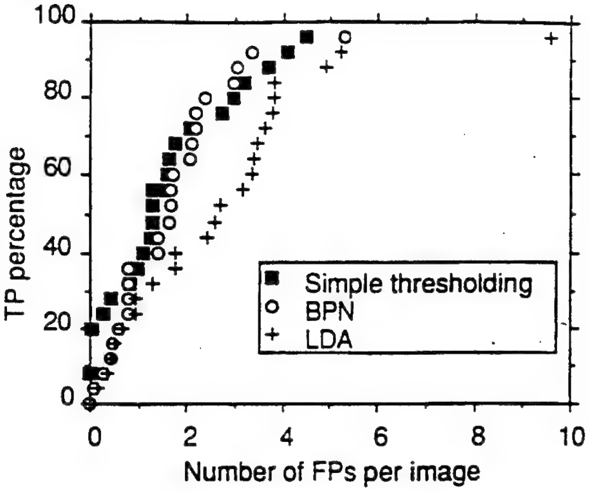


Fig. 10. Plot of the tradeoff between TP and FP detections for thresholding, LDA and BPN classification.

The morphological features from each of the 461 split objects were calculated. The sets of features were then used to classify the objects as potential masses using the thresholding, LDA, and BPN classification schemes. To perform this training classification, the feature sets were input into the classifiers with known desired output for each object and the classifiers were trained to provide the best classification. Fig. 10 summarizes the trade-off between the TP fraction and the number of FP detections for each of the three classification methods, using both the mass and nonmass training features.

V. Discussion

The first stage of DWCE segmentation is applied globally to the entire breast image. Its primary function is to define a set of local regions likely to contain the true mass objects. In the present study, the first stage provided this capability by detecting 24 of the 25 true masses in our data set. Splitting was not applied to the detected objects in the first stage because the detected regions were usually much smaller than the true structures seen on the original mammogram. In this study, the average area of the first stage objects was 54.5 pixels. These smaller objects produced less region merging. Therefore, only a simple thresholding classification was used to reduce the initial number of regions. For comparison purposes, LDA and BPN classification were also applied following the first DWCE stage. The best classifier produced over 10 FP's per image at a 90% TP detection rate. This result is significantly larger than the results from the second-stage classification and highlights the need for a second filtering step.

The second DWCE filter is the main detection stage of the segmentation scheme. The enhancement is applied locally to each object ROI. This allows the filter to adapt to the intensity distribution within each ROI, thereby reducing the effects of intensity variation across the full mammogram as seen by the first filtering stage. The improved detection leads to larger objects having an average area of 70.2 pixels. The increased object size can be attributed to both more precise edge information (i.e., better edge localization) and region merging associated with the segmentation. Because there was significant region merging in the second stage, splitting was applied to all the potential mass objects. Object splitting allows the merged regions to be separated without losing any of the detected masses. However, it can introduce false edge locations in the detected objects, lessening the effectiveness of the classification stage. Alternatively, merging can be reduced by applying a more stringent DWCE filter. Note that stringent DWCE filters can be created by increasing the intensity thresholds of $K_M(\cdot)$ (minimum value where $K_M(\cdot) \geq 1.0$) and $K_{NL}(\cdot)$ (minimum value where $K_{NL}(\cdot) = 1.0$). However, it was found that there was a tradeoff between region merging and the number of missed masses. The current combination of DWCE filter parameters and region splitting was found to provide the best mass detection capability and produced a minimum number of regions compared with other combinations evaluated. The flexible form of the DWCE filter leaves open the possibility that further optimization of the detection can be explored with a completely different set of filter parameters. Evaluation of different DWCE filters with different functions for $K_M(\cdot)$ and $K_{NL}(\cdot)$ will be pursued in future studies. In this preliminary study, our goal is to demonstrate the feasibility of using DWCE filters for the detection of breast masses.

The final classification stage was used to further reduce the number of FP regions detected. The training results for the thresholding, LDA, and BPN classifiers were compared using the 461 objects produced by the second DWCE stage. The number of FP's was initially reduced from 18.4 regions per image to only 4.5 regions without increasing the number of missed masses. For a 90% TP detection rate, the FP's can be reduced to 3.0 per image. Fig. 10 shows that the thresholding method and the BPN classifier provided comparable results with the BPN slightly better when a few additional misses can be tolerated. On the other hand, the LDA classifier consistently produced a slightly larger FP rate for a given TP rate compared with both the thresholding and BPN classifiers. However, the only significant difference between LDA and the other two classifiers is seen at the 96% TP detection rate. Again, the training results in Fig. 10 are based on only five morphological features. Additional features may improve individual classifier performance. Further reduction in the FP's may also be obtained using more sophisticated tissue classification algorithms either independently or in conjunction with the morphological classifiers [26], [27].

A TP detection rate of 96% with only 4.5 FP per image shows that the DWCE can be used to detect breast masses on digitized mammograms. Its main advantage is that it adapts the enhancement to the local density or background in the image. This enables subtle as well as obvious masses superimposed on structured background to be detected. Since the DWCE provides high frequency edge information, morphological features based on object boundaries can be used in combination with a classification scheme to reduce the number of detected regions. The edge information, however, is not complete because of region merging and the subsequent splitting operation which introduces further errors in edge localization. The edge locations are also affected by the DWCE filter parameters as seen in Fig. 9. The current DWCE implementation produces conservative estimates for the true edges of the objects. In other words, the estimated edges fall within the true boundary for isolated breast structures.

VI. Conclusion

The results of the DWCE segmentation indicates that it is a viable option for automated mass detection in mammography. It effectively segmented the digitized mammograms into a small number of potential breast masses without a significant loss in the number of true masses. In this preliminary study, the thresholding, LDA and BPN morphological feature classifiers were evaluated using a limited training data set. The initial results indicate that nonlinear combinations of the features are slightly more effective for FP reduction. Further studies are currently under way using a larger set of images. This larger image set will be used to determine if additional morphological features are necessary and to determine if the BPN classifier is truly the best choice. The utility of the currently trained DWCE segmentation method (i.e., structure, filter and classification parameters) will also be evaluated using a unique subset of test images from the new database. A study involving a set of extracted ROI's from the original high resolution mammograms based on the detected DWCE objects is also being conducted. Its purpose is to investigate if more sophisticated feature extraction and tissue classification schemes can further reduce the number of FP detections and determine the malignancy of each detected mass.

References

[1] C. C. Boring, T. S. Squires, T. Tong, and S. Montgomery, "Cancer statistics, 1994," *CA—A Cancer J. Clinicians*, vol. 44, no. 1, pp. 7–26, Jan. 1994.

[2] S. Shapiro, W. Venet, P. Strax, L. Venet, and R. Roeser, "Ten-to-fourteen-year effect of screening on breast cancer mortality," *JNCI*, vol. 69, pp. 349, 1982.

[3] R. G. Lester, "The contributions of radiology to the diagnosis, management, and cure of breast cancer," *Radio.*, vol. 151, p. 1, 1984.

[4] M. Moskowitz, "Benefit and risk," in *Breast Cancer Detection: Mammography and Other Methods in Breast Imaging*, 2nd ed., L. W. Bassett and R. H. Gold, Eds. New York: Grune and Stratton, 1987.

[5] L. Tabár and P. B. Dean, "The control of breast cancer through mammography screening: What is the evidence," *Radiol. Clin. N. Amer.*, vol. 25, no. 5, pp. 993–1005, Sept. 1987.

[6] J. Caseldine, R. Blamey, E. Roebuck, and C. Elston, *Breast Disease for Radiographers*. Toronto, Canada: Wright, 1988.

[7] J. E. Martin, M. Moskowitz, and J. R. Milbrath, "Breast cancer missed by mammography," *Am. J. Roentgeno.*, vol. 132, pp. 737–739, May 1979.

[8] R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiol.*, vol. 184, pp. 613–617, Sept. 1992.

[9] M. G. Wallis, M. T. Walsh, and J. R. Lee, "A review of false negative mammography in a symptomatic population," *Clin. Radiol.*, vol. 44, pp. 13–15, 1991.

[10] F. Shtern, C. Stelling, B. Goldberg, and R. Hawkins, "Novel technologies in breast imaging: National Cancer Institute perspective," in Second Post-Graduate Course Syllabus, Soc. of Breast Imaging, Orlando, FL, May 1995, pp. 153–156.

[11] H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms," *Investigative Radiol.*, vol. 25, no. 10, pp. 1102–1110, Oct. 1990.

[12] W. P. Kegelmeyer Jr., "Computer detection of stellate lesions in mammograms," in *Proc. SPIE Biomed. Image Processing*, 1992, vol. 1660, pp. 446–454.

[13] S. M. Lai. X. Li. and W. F. Bischof. "On techniques for detecting circumscribed masses in mammograms." *IEEE Trans. Med. Imag.*, vol. 8. no. 4. pp. 377–386. Dec. 1989.

[14] W. Qian. L. P. Clarke. M. Kallergi. and R. A. Clark. "Tree-structured nonlinear filters in digital mammography. *IEEE Trans. Med. Imag.*, vol. 13. no. 1. pp. 25–36. Mar. 1994.

[15] D. Brzakovic. X. M. Luo. and P. Bzrakovic. "An approach to automated detection of tumors in mammography." *IEEE Trans. Med. Imag.*, vol. 9. no. 3. pp. 233–241. Sept. 1990.

[16] F. F. Yin. M. L. Giger. K. Doi. C. E. Metz. R. A. Vyborny. and C. J. Schmidt. "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images." *Med. Phys.*, vol. 18. no. 5, pp. 955–963. Sept. 1991.

[17] T. K. Lau and W. F. Bischof. "Automated detection of breast tumors using the asymmetry approach." *Comput. Biomed. Res.*, vol. 24. pp. 273–295. 1991.

[18] W. P. Kegelmeyer Jr., J. M. Pruneda, P. D. Bourland. A. Hillis. M. W. Riggs. and M. L. Nipper. "Computer-aided mammographic screening for spiculated lesions." *Radiol.*, vol. 191. no. 2. pp. 331–337, May 1994.

[19] T. Peli and J. S. Lim. "Adaptive filtering for image enhancement." *Opt. Eng.*, vol. 21. no. 1, pp. 108–112. 1982.

[20] W. K. Pratt. *Digital Image Processing*. New York: Wiley. 1978.

[21] D. Marr and E. Hildreth. "Theory of edge detection." in *Proc. Royal Soc. London Ser. B Bio. Sci.*, vol. 207. 1980 pp. 187–217.

[22] W. H. H. J. Lunscher and M. P. Beddoes. "Optimal edge detector design i: Parameter selection and noise effects." *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8. no. 2. pp. 154–176. Mar. 1986.

[23] ——, "Optimal edge detection design ii: Coefficient quantization." *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8. no. 2. pp. 178–187. Mar. 1986.

[24] P. A. Lachenbruch. *Discriminant Analysis* New York: Hafner. 1975.

[25] J. A. Freeman and D. M. Skapura. *Neural Networks: Algorithms. Applications and Programming Techniques*. Reading. MA: Addison–Wesley. 1991.

[26] H. P. Chan. D. Wei. M. A. Helvie. B. Sahiner. D. D. Adler. M. M. Goodsitt. and N. Petrick. "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space." *Phys. Med. Bio.*, vol. 40. pp. 857–876. 1995.

[27] D. Wei. H. P. Chan. M. A. Helvie. B. Sahiner. N. Petrick. D. D. Adler. and M. M. Goodshitt. "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis." *Med. Phys.*, May 1995.

Image compression in digital mammography: Effects on computerized detection of subtle microcalcifications

Heang-Ping Chan,[a] Shih-Chung B. Lo,[b] Loren T. Niklason,[c] Debra M. Ikeda,[d] and Kwok Leung Lam[e]

Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 13 June 1995; resubmitted 14 March 1996; accepted for publication 29 May 1996)

Our previous receiver operating characteristic (ROC) study indicated that the detection accuracy of microcalcifications by radiologists is significantly reduced if mammograms are digitized at 0.1 mm×0.1 mm. Our recent study also showed that detection accuracy by computer decreases as the pixel size increases from 0.035 mm×0.035 mm. It is evident that very large matrix sizes have to be used for digitizing mammograms in order to preserve the information in the image. Efficient compression techniques will be needed to facilitate communication and archiving of digital mammograms. In this study, we evaluated two compression techniques: full frame discrete cosine transform (DCT) with entropy coding and Laplacian pyramid hierarchical coding (LPHC). The dependence of their efficiency on the compression parameters was investigated. The techniques were compared in terms of the trade-off between the bit rate and the detection accuracy of subtle microcalcifications by an automated detection algorithm. The mean-square errors in the reconstructed images were determined and the visual quality of the error images was examined. It was found that with the LPHC method, the highest compression ratio achieved without a significant degradation in the detectability was 3.6:1. The full frame DCT method with entropy coding provided a higher compression efficiency of 9.6:1 at comparable detection accuracy. The mean-square errors did not correlate with the detection accuracy of the microcalcifications. This study demonstrated the importance of determining the quality of the decompressed images by the specific requirements of the task for which the decompressed images are to be used. Further investigation is needed for selection of optimal compression technique for digital mammograms. © *1996 American Association of Physicists in Medicine.*

Key words: mammography, digital, microcalcifications, image compression, computer-aided diagnosis

I. INTRODUCTION

X-ray mammography is the most effective method in detection of early breast cancers.[1] Because of the stringent requirements for imaging of subtle lesions, the image recording systems for mammography have to provide very high spatial resolution and high contrast sensitivity. At present, screen-film systems specially designed for mammography are the only recording medium that can provide the image quality needed. However, because of the advancement in digital imaging technology, digital mammography is becoming a realistic goal. Digital mammography offers the advantages of electronic transmission, consultation, and archiving as well as image enhancement and computer-aided diagnosis.[2] These will potentially make mammography more widely accessible, reduce the cost, and improve the diagnostic accuracy of mammography.

Our previous receiver operating characteristic (ROC) study indicated that the detection accuracy of microcalcifications by radiologists is significantly reduced if mammograms are digitized at 0.1 mm×0.1 mm.[3] Our recent study also showed that detection accuracy by computer decreases as the pixel size increases from 0.035 mm×0.035 mm.[4] It is evident that very high resolution digitization has to be used for mammograms in order to preserve the information in the image. A 18 cm×24 cm mammogram digitized at 0.05 mm ×0.05 mm results in a matrix size of about 4000×5000. A four-view study thus will provide 160 megabytes of digital data. The transmission and archiving of such a large amount of data is therefore one of the important considerations in implementation of digital mammography. An efficient data compression scheme that can reduce the amount of data without degradation of the image quality for human and machine interpretation will alleviate these problems.

Much effort has been devoted to evaluate compression methods for radiological images. Most studies so far applied to digital chest radiography[5–10] because it is the most commonly performed procedure in medical imaging, and some direct digital imaging systems for chest radiography are already available. Recently several investigators have extended their studies to general radiological images, including computed tomography (CT), and magnetic resonance (MR) images based on DCT[11–13] and wavelet-type decomposition methods.[14–16] Some preliminary studies have also been performed for digitized mammographic images.[17–19]

In this study, we explored some of the issues involved in compression of mammographic images for applications in computerized detection of microcalcifications.[20] Because primary digital mammography systems are not yet available,

0094-2405/96/23(8)/1325/12/$10.00

digital mammograms in this study were obtained by digitization of screen-film mammograms. We selected two image compression techniques, the Laplacian pyramid hierarchical coding (LPHC)[21] and the discrete cosine transform (DCT), with full frame entropy coding (FFEC)[5,6] for processing of mammograms. The LPHC technique was chosen because of its similarity to the difference-image technique that we used for enhancement of microcalcifications in the automated detection algorithm. The DCT technique was commonly used for an irreversible image compression and FFEC was developed to eliminate the block artifacts and improve compression efficiency. The DCT-FFEC technique was further implemented using bit plane splitting to improve the preservation of detailed information.[7] We compared the compression efficiency of these techniques for digitized mammograms. The fidelity of the information in the reconstructed image was evaluated by the detectability of the microcalcifications by an automated computer program. The results were compared with the mean square error (MSE), which was a commonly used indicator of information loss in image compression.

II. MATERIALS AND METHODS

A. Data set of digital mammograms

Twenty-five mammograms were selected from patient files from the Department of Radiology at the University of Michigan. All mammograms were acquired with American College of Radiology accredited machines and recorded with Kodak Min R/Min $R-E$ screen-film systems. Each mammogram contained a cluster of subtle microcalcifications, the presence of which had been verified by biopsy. The mammograms were digitized with a high-resolution laser scanner at a pixel size of 35 μm×35 μm and 12-bit gray levels. The digitizer logarithmically amplified the transmitted light through the film before digitization. The scanner was calibrated such that the gray levels were linearly proportional to optical density (OD) in the range of about 0.1–2.8 OD. The optical density range of the scanner was 0–3.5 OD.

Because of the computational requirement for processing the entire breast image that could be greater than 4000×5000 pixels, we manually extracted an ROI of 1024×1024 pixels, which contained the cluster of microcalcifications from each digitized image. The ROIs were used as input images in the compression and detection studies. To establish a "truth" file for the microcalcifications, the coordinate of each individual microcalcification in an ROI was identified manually with a cursor on a display workstation. The locations of the microcalcifications were verified by visually compared with those on the film mammograms using a magnifier. The coordinates were stored in a "truth" file and used for scoring the detection accuracy by the automated procedure, as discussed below. The total number of microcalcifications in the 25 ROIs was 293.

B. Laplacian pyramid hierarchical coding (LPHC)

The LPHC is a noncausal image coding method that decomposes an image into a low-pass image and a sequence of sub-band images, each of which is reduced in spatial resolution by a factor of 2, thereby forming a pyramidal hierarchical structure.[21] The LPHC technique implemented for this study is described in Appendix A. This compression method is developed for progressive image transmission. The low resolution version (the top level in the Gaussian pyramid) of the image is transmitted first to provide an early impression of the image content, progressively higher resolution images are subsequently transmitted to provide greater details. Transmission can be terminated as soon as sufficient image information is received. If the low level images are not needed and therefore not transmitted, the number of bits per image in the transmission is greatly reduced. Furthermore, the image size of the top level Gaussian pyramid image is small and the entropy of the Laplacian pyramid images is low because of the removal of the pixel-to-pixel correlation, the coding of the decomposed images can be more efficient than that of the original image. Further image compression can be achieved by reducing the quantization levels of the pixel values of the Laplacian pyramid images. For the purpose of this study, we will investigate the effects of the image reconstruction levels and the quantization levels of the Laplacian pyramid images on detection accuracy by the computer program.

C. Discrete cosine transform-full frame entropy coding (DCT-FFEC)

Block-DCT techniques are commonly used for compression of continuous-tone digital images. DCT can effectively localize most of the image information (energy) in a small area in the spatial frequency domain. However, the division of the image into small blocks for DCT often introduces blocky artifacts to the reconstructed images when high compression ratios are desired.

The full frame DCT technique transforms the entire image in one block. It not only eliminates the blocky artifacts, but also provides the advantage that the large-size DCT can localize the image information in a relatively smaller bandwidth than the small-size DCT. The coefficients in the full frame DCT matrix can be quantized with linear or nonlinear methods and then encoded by various coding techniques. For example, with a full frame bit allocation (FFBA) technique,[5] a bit-allocation table based on the characteristics of the transformed image and the desired compression ratio is produced. The table indicates the number of bits designated for a specific coefficient or groups of coefficients. The quantized coefficients are then packed into the bit space indicated in the table.

More recently, an entropy coding scheme that does not require a bit allocation table was developed for full frame DCT. For chest radiographs and CT images, the FFEC method was found to be more efficient than FFBA, in that it could produce a lower degree of MSE at a given compression ratio or increased the compression efficiency with the same MSE.[6] These studies also indicated that a bit splitting-remapping method was useful in preventing errors in encoding for the most significant bits and to lessen edge artifacts caused by compression and decompression.[7] We therefore investigated the FFEC technique with and without bit splitting remapping for image compression in computer-aided diagnosis (CAD) applications. A description of the FFEC technique implemented for this study is given in Appendix B.

D. Mean-square error (MSE)

A commonly used indicator of information loss in a compression–decompression scheme is the MSE between the original image, $g(x,y)$, and the decompressed image, $g_r(x,y)$, as given by

$$\text{MSE}=\frac{1}{N_p}\sum_{\forall x}\sum_{\forall y}[g(x,y)-g_r(x,y)]^2, \tag{1}$$

where N_p is the total number of pixels in the image. The MSE is a global measurement of distortion of the image by a lossy compression technique. The use of MSE as an indicator of information loss in our computerized detection of microcalcifications on compressed–decompressed mammograms was evaluated in this study.

E. Computerized detection of microcalcifications

We have described our CAD algorithm for detection of microcalcifications in detail previously.[4,20,22,23] Briefly, there are three major steps in the algorithm: preprocessing, segmentation, and classification. In the preprocessing step, the input digital mammogram is processed with a signal-enhancement filter and a signal-suppression filter. The difference of these two filtered images results in an image in which the structured background is suppressed and the signal-to-noise ratio (SNR) of the microcalcifications is enhanced. This is also referred to as a difference-image technique. In the segmentation step, the program determines the gray level histogram of the processed image within the breast region. A gray level thresholding technique is used to locate potential signal sites above a global threshold. The threshold is changed iteratively until the number of sites obtained falls within the chosen input maximum and minimum numbers. At each potential site, a locally adaptive gray level thresholding technique in combination with region growing is performed to segment the connected pixels above a local threshold, which is calculated as the product of the local root-mean-square (rms) noise and an input SNR threshold. The characteristics of a segmented signal such as the size, contrast, SNR, and its location, are determined.

In the classification step, the computer program performs three tests to distinguish signals from noise or artifacts. A lower bound is imposed on the size to exclude signals below a certain size, which are likely to be noise, and an upper bound is set to exclude signals greater than a certain size, which are likely to be large benign calcifications. A contrast upper bound is also set to exclude potential signals that have a contrast higher than an input number of standard deviations above the average contrast of all potential signals found with local thresholding. This criterion excludes the very high-contrast signals that are likely to be artifacts and large benign calcifications. A regional clustering procedure is then applied to the remaining signals; a signal is kept if the number of signals found within a neighborhood of a chosen input diameter around that signal is greater than an input minimum number. The remaining signals that are not found to be in the neighborhood of any potential clusters will be considered isolated noise points or isolated calcifications and excluded. This clustering criterion is useful for reducing false positives because true microcalcifications of clinical interest always appear in clusters on mammograms. The specific parameters used in each step have been described previously.[4]

In this study, a signal-enhancement filter of 2×2 kernel of constant weights and a signal-suppression filter, which was a box-rim filter with a 20×20 kernel of constant weights around the rim and a 12×12 central area of 0 weights,[4] were used for preprocessing of the decompressed images obtained with the DCT-FFEC techniques. The sum of the weights was normalized to unity in each of the filters. For the images compressed with the LPHC technique, we made use of the Laplacian pyramid images to directly generate the difference image. The decoded Laplacian pyramid images at all levels were expanded to the original image size, summed together, and convolved with the 5×5 kernel, $w(m,n)$, defined in Appendix A. The resulting bandpass image was used as the difference image. This was equivalent to using the decompressed image as the signal-enhanced image and the Nth-level Gaussian pyramid image expanded N times to the original image size as the signal-suppressed image, and convolving the difference between the two images with the 5×5 kernel.

F. Analysis of detection accuracy

Afte. passing the size, contrast, and the regional clustering criterion, the detected individual microcalcifications would be compared with the "truth" file of the input image. The numbers of true-positive (TP) and false-positive (FP) microcalcifications were scored. A detected signal was scored as a TP microcalcification if it was within 0.35 mm from a true microcalcification in the "truth" file. Once a true microcalcification was matched to a detected microcalcification, it would be eliminated from further matching. Any detected microcalcifications that did not match to a true microcalcification were scored as FPs. The trade-off between the TP and FP detection rates by the computer program was evaluated by the free-response receiver operating characteristic (FROC) analysis[24] by varying the input SNR threshold. A low SNR threshold corresponded to a lax criterion with a large number of FPs. A high SNR threshold corresponded to a stringent criterion with a small number of FPs and a loss in TPs. In this study, the FP rate was expressed as the number of FPs per unit area of the ROI image in order to reduce its dependence on the image size.[4] The information content of the reconstructed images was then evaluated by comparison of the FROC curves, which indicated the detection accuracy of the computer program.

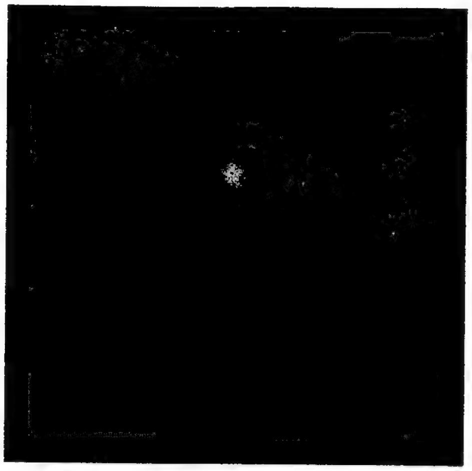


FIG. 1. An original ROI image with a cluster of microcalcifications used in the dataset of this study. A subtle cluster of about ten microcalcifications is located near the center of the ROI.

III. RESULTS

Figure 1 is an example of a 1024×1024-pixel ROI from a mammogram digitized to 12 bits. A subtle cluster of about ten microcalcifications is located near the center of the ROI. Three levels of the Laplacian images of the ROI at 12 bits, as well as the corresponding Laplacian images quantized to eight bits, are shown in Fig. 2(a). It can be seen that the frequency range of the information in the Laplacian images decreased with increasing levels on the pyramid. The gray level histograms of the two level-0 Laplacian images were plotted in Fig. 2(b), which illustrates the low entropy (defined in Appendix A) in a Laplacian pyramid image and the further reduction of entropy by requantization. To demonstrate visually the effect of requantization on image fidelity, we reconstructed the ROI from the three levels of eight-bit Laplacian pyramid images and the top level of the Gaussian pyramid, as shown in the flow diagram of LPHC in Appendix A. The error image between the original 12-bit image in Fig. 1 and the reconstructed image is shown in Fig. 3(a).

For the LPHC method, we have attempted three different ways to eliminate the LSBs in the Laplacian images. In the first method, each pixel value was divided by 2^l (l is the number of LSBs to be eliminated) and rounded off to the nearest integer. The pixel value was multiplied by 2^l during reconstruction. The second method was similar to the first, except that the quotient was truncated to an integer. The third method simply set the LSB to 0's by a bitwise AND operation with a bit plane mask and was the most efficient in terms of computational speed among the three. The third method yielded an image different from the second method because the Laplacian image was a difference image that contained negative integers. The truncation method reduced the absolute values of both the positive and negative integers, whereas the bit masking method shifted both the positive and negative integers to lower values (i.e., more negative for negative integers). The error images between the original and the reconstructed images with eight-bit quantization using the three different bit reduction methods are shown in Figs. 3(a)–3(c), respectively. The round-off and the bit masking methods for bit reduction resulted in similar error images, but the error image from the truncation method had obvious noise patterns. As described below, the MSE of the truncation method was much larger than those of the other two methods. The FROC curves for the LPHC techniques shown in Figs. 4 and 6 were obtained with the bit masking method because of its computational speed and its similarities to the round-off method.

With the LPHC method, both the compression ratio and the reconstruction accuracy depend on the number of levels on the pyramid that the image is decomposed and reconstructed. We first investigated the dependence of the detection accuracy of the computer algorithm on the pyramid level. Figure 4 shows the FROC curves for two to four Gaussian pyramid levels of decomposition and reconstruction. The Laplacian images were quantized to eight bits in all cases. The images decomposed into three levels and reconstructed provided higher accuracy than those of two and four levels. Applying a paired t test to the TP rates at corresponding FP rates and pooled over the range of FPs between about 0.1 and 1 FP per cm^2, as discussed previously[4] and in Sec. IV, we found that the TP rates for the three-level decomposed images are significantly higher, with a two-tailed p value of less than 0.001, than those for the two- or four-level decomposed images.

The dependence of the detection accuracy on reconstruction level was also examined. The images were decomposed to three levels and the Laplacian pyramid images were maintained at 12 bits. The images were then reconstructed to the second level (image size of 256×256 pixels), to the first level (image size of 512×512 pixels), and to the original level (image size of 1024×1024 pixels). Because the first and second level images were already low-pass filtered, the 5×5 kernel was not applied to the difference images in the detection process in order to avoid further reduction in the spatial resolution. The detection results were plotted in Fig. 5. The detection accuracy decreased drastically if the images were not reconstructed to the original zeroth level. The decreases in the TP rates are statistically significant from the original level to the first level ($p<0.01$) and from the original to the second level ($p<0.001$).

Based on these results, images decomposed to the third level and reconstructed to the original level provided the highest detection accuracy. The following studies of the dependence of the detection accuracy on the bit depth of quantization were performed under this condition. The quantization of the Laplacian pyramid images was varied from six bits to nine bits. The FROC curves for these quantization bit depths were plotted in Fig. 6, along with the FROC curve for the original 12-bit images. There are no statistically significant differences among the curves for the 12-bit to 8-bit images ($p>0.05$). As the bit depth decreased further to seven bits, the reduction in the TP rates from those of the 12-bit images in the range of FP between 0.1 and 1 per cm^2 became statistically significant at $p<0.003$. The corresponding average bit rate for each of the conditions is shown in the figure legends. It indicates that, at eight-bit quantization of the La-

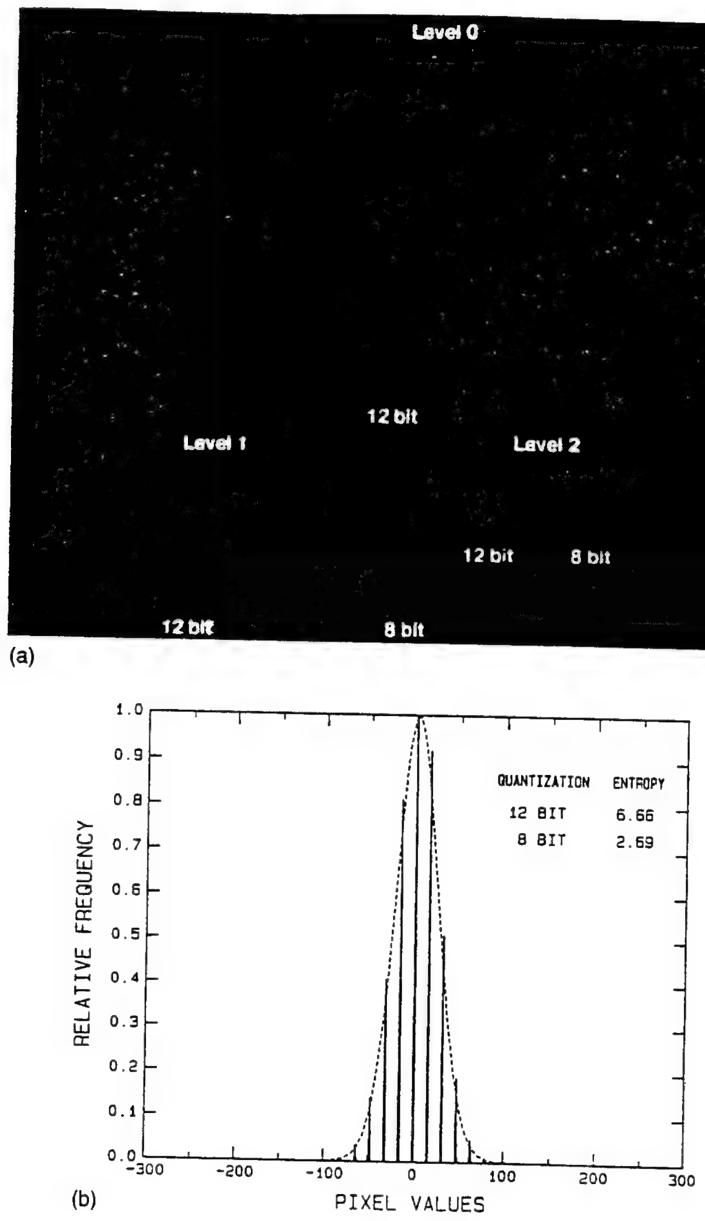


FIG. 2. (a) Three levels of Laplacian pyramid images of the ROI in Fig. 1. At each level the original 12-bit image is compared to the 8-bit quantized image. Note that the Laplacian pyramid images contain both positive and negative pixel values. For display purposes, a constant was added to every pixel of each image to shift them to the same positive mean level. (b) The gray level histograms of the level-0 Laplacian images at 12-bit (dashed curve) and at 8-bit (solid curve) gray level resolution.

placian images, the images can be compressed to an average of 3.28 bits/pixel without loss of the detectability of the microcalcifications by the computer algorithm.

With the DCT-FFEC approach, we evaluated three conditions, splitting into MSB of 3 and LSB of 9, splitting into MSB of 4 and LSB of 8, and without splitting. The detection accuracy in the reconstructed images for these conditions was compared in Figs. 7(a)–7(c). The parameters used for the compression schemes are listed in Table I in Appendix B. With splitting, the detection accuracy for the microcalcifications was similar to that in the original images at entropy coding ranges of seven to two bits and six to two bits ($p > 0.05$). The coding ranges of six to two bits resulted in an average bit rate of 1.25 for the three- and nine-bit splitting and 1.82 for the four- and eight-bit splitting. The detectability dropped significantly ($p < 0.001$) as the range reduced to five to two bits for both splitting schemes. The error image for the three- and nine-bit splitting with the entropy coding range of six to two bits is shown in Fig. 3(d). Without splitting, the detection accuracy for the entropy coding ranges of eight to two bits and seven to two bits was comparable to that of the original images ($p > 0.05$). The average bit rate for the seven to two bits coding was 1.49. The detectability was significantly lower for the coding range of six to two bits ($p < 0.001$).

The MSE and the bit rate for each of the compression

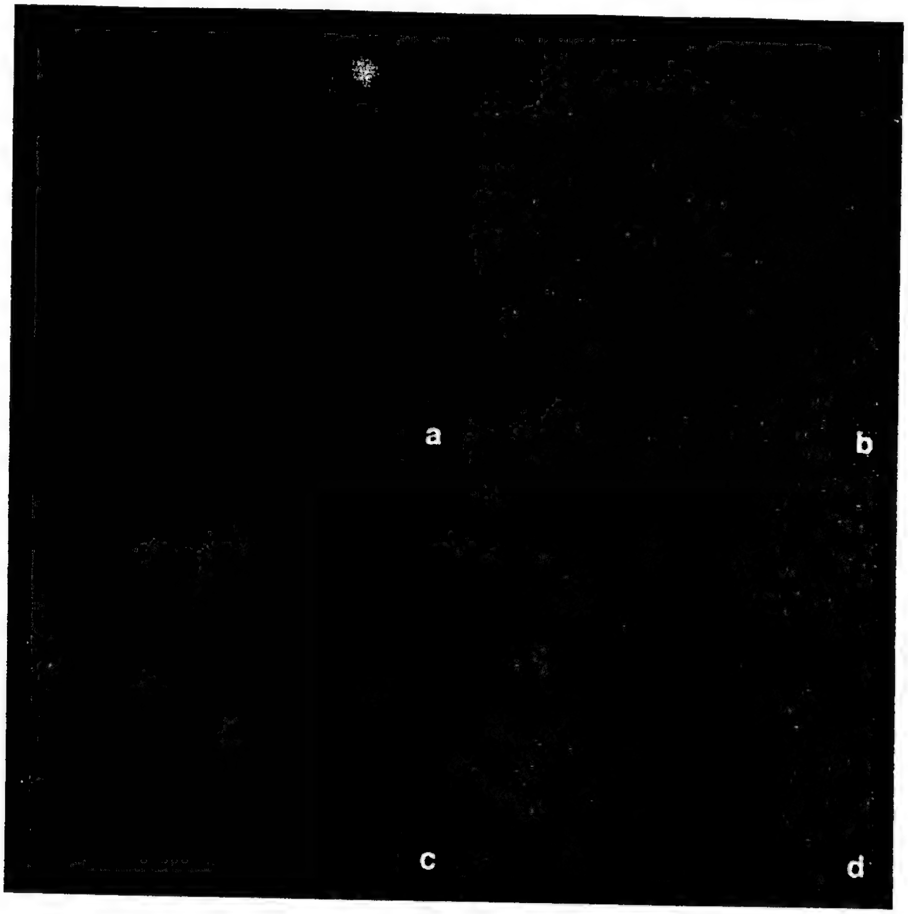


FIG. 3. The error images between the original image in Fig. 1 and the reconstructed images. Note that, for display purposes, a constant was added to every pixel of the error images to shift them to the same positive mean level. (a)–(c). The reconstructed images were obtained with the LPHC compression–decompression scheme. The image was decomposed to the second level of the Laplacian pyramid ($N=3$), quantized to eight bits, and reconstructed to the zeroth level. The eight-bit quantization was obtained by (a) dividing by 2^4 and rounding off to the nearest integers, (b) dividing by 2^4 and truncated to integer, and (c) setting the four LSBs to 0 using a bitwise AND operation with a bit plane mask. (d) The reconstructed image was obtained with the DCT-FFEC technique, three MSB and nine LSB splitting, entropy coding range of six to two bits.

schemes and conditions were averaged over the set of 25 images. The results were plotted on a semilog scale in Fig. 8. Each of the solid curves shows the results for a DCT-FFEC scheme with a different bit splitting parameter. The range of bits for entropy coding was varied along each curve. The results for the LPHC method were plotted as dashed curves. The lowest average bit rate that a scheme could achieve at which the detectability of the microcalcifications by computer was comparable to that of the original images was identified by an arrow for each of the compression schemes. The logarithm of the MSE appeared to be inversely proportional to the bit rate for most compression schemes. The curves for the DCT-FFEC schemes without or with splitting were comparable. The lowest average bit rate achieved by the DCT-FFEC method without a degradation in the detection accuracy is 1.25, corresponding to a compression ratio of about 9.6:1 in comparison to a 12-bit images without compression.

For the LPHC technique, the curve for each bit elimination method was plotted in Fig. 8 for linear quantization from six to nine bits. It can be seen that the round-off and the bit masking methods yielded similar MSE and bit rates for

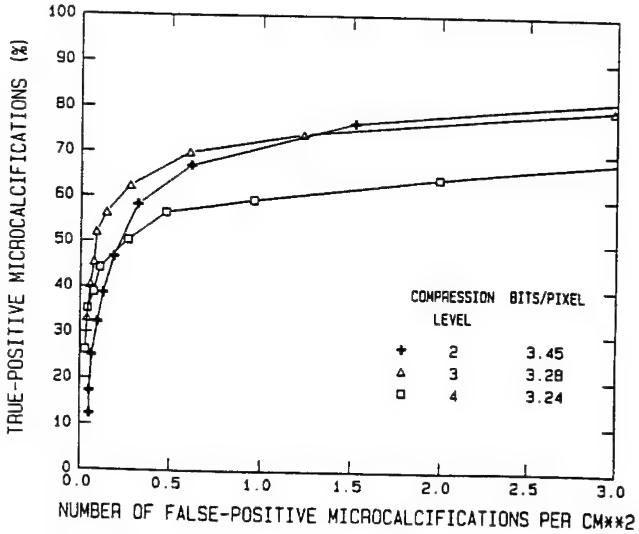


FIG. 4. The FROC curves for images decomposed to the Nth level (refer to the flow diagram in Fig. 9) and reconstructed to the zeroth level; crosses: $N=2$, triangles: $N=3$, and squares: $N=4$. The Laplacian pyramid images were quantized to eight bits.

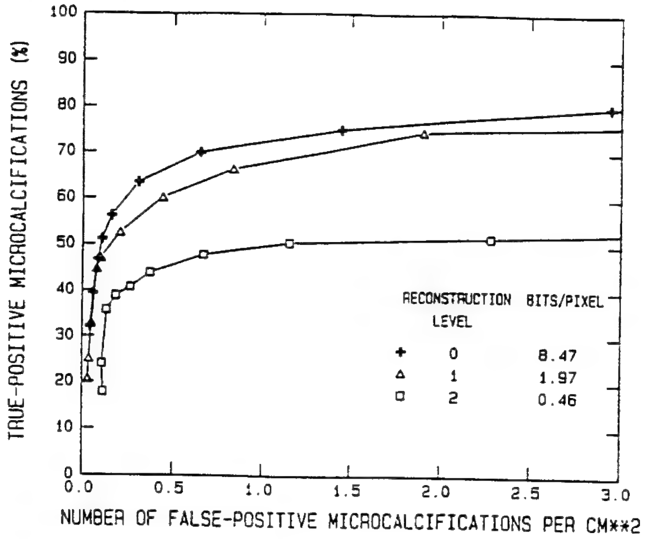


FIG. 5. The FROC curves for images decomposed to the $N=3$ level and reconstructed to the zeroth level (crosses), the first level (triangles), and the second level (squares). The Laplacian pyramid images were used at 12 bits without requantization.

seven to nine bit quantization. For the truncation method, the MSE was much higher than the other two methods for a given number of quantization bits. This is consistent with the visual appearance of the error image shown in Fig. 3(b). The entropy of the Laplacian images from the truncation method was lower than those from the other methods because a larger number of pixel values were truncated to zero from both the positive and the negative pixel values. Although it appeared that the detectability of the eight-bit truncated images did not decrease significantly compared to that of the original images, the detectability dropped more rapidly than the other two methods when the Laplacian images were further truncated to seven bits and six bits. Using the bit masking or round-off method, the lowest bit rate achieved without a degradation in the detectability of the microcalcifications by the computer was about 3.3, which corresponded to a

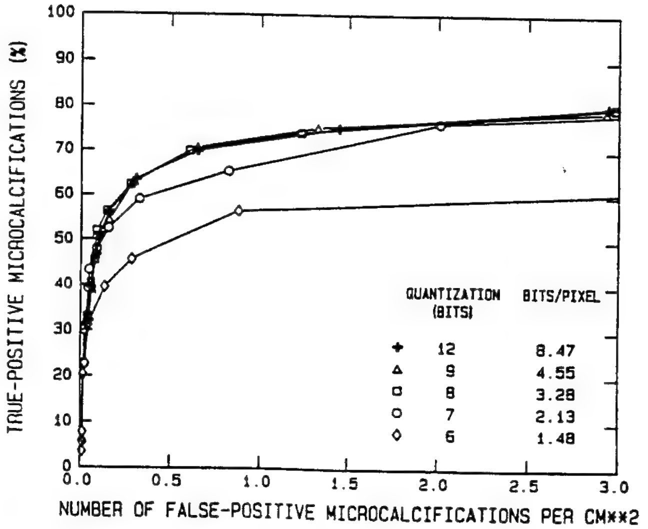


FIG. 6. The FROC curves for images decomposed to the $N=3$ level and reconstructed to the zeroth level. The Laplacian pyramid images were quantized to six to nine bits. The bit rate of 8.47 at 12 bits was calculated for a lossless compression with the LPHC technique.

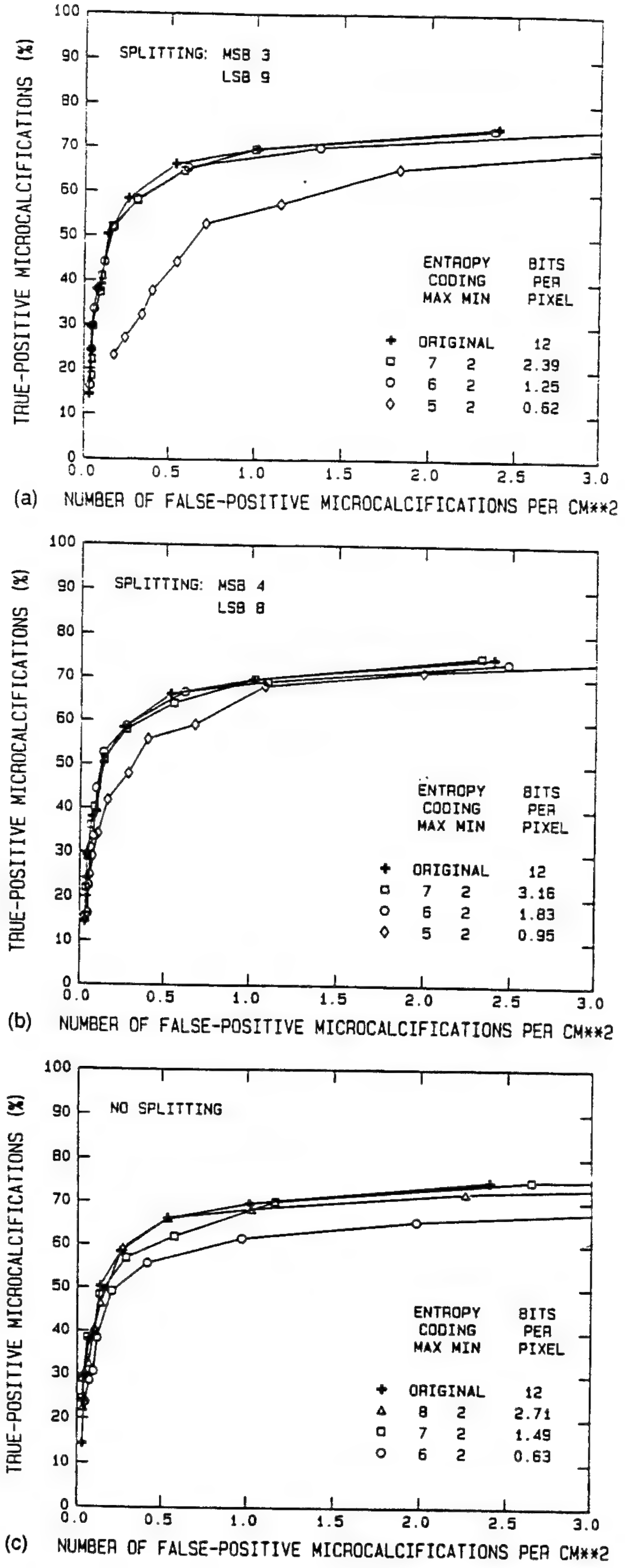


FIG. 7. The FROC curves for images compressed and decompressed with the DCT-FFEC technique, (a) with splitting at MSB=3 and LSB=9, (b) with splitting at MSB=4 and LSB=8, and (c) without splitting.

compression ratio of 3.6:1 in comparison to a 12-bit image without compression.

IV. DISCUSSION

The results of this study indicate that the DCT-FFEC method can provide a higher compression ratio than the LPHC method. The DCT-FFEC method with bit splitting of three MSB and nine LSB and entropy coding of six to two bits can achieve a compression ratio of 9.6:1 without significant degradation in the detectability of microcalcifications by a computer algorithm. On the other hand, with the LPHC method and the range of parameters studied, the compression ratio is only about 3.6:1 if the detectability of subtle microcalcifications has to be preserved. The DCT-FFEC method is thus about three times more efficient than the LPHC method if the detectability of microcalcifications by computer is used as the criterion of image fidelity. Although computer vision can be very different from human vision and the results cannot be simply generalized to image compression to be used for human readers, our results indicate that the DCT-FFEC method can retain high-frequency information, such as that of the microcalcifications better than the LPHC method.

Examples of the error images obtained from subtracting the decompressed image from the original image for the different compression schemes with the selected parameters are shown in Figs. 3(a)–3(d). It can be seen that the error image from the DCT-FFEC technique [Fig. 3(d)] appears to be more random and contain higher frequencies than those from the LPHC technique. The error image from the LPHC technique with truncation [Fig. 3(b)] contained visible structures of the mammogram. This problem can be avoided by a constant shift of all pixel values of the Laplacian pyramid images to positive integers before division and truncation. The resulting image and detectability of microcalcification will then be similar to the round-off and bit-masking methods.

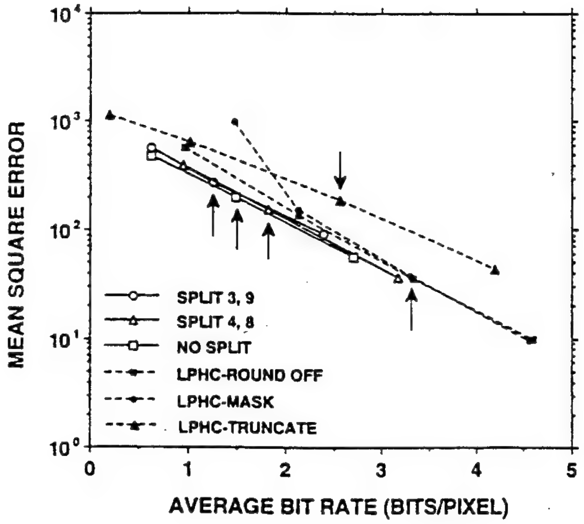


FIG. 8. The relationship between the mean square error (MSE) and the average bit rate for the various compression techniques. The solid curves are for DCT-FFEC techniques and the dashed curves are for the LPHC techniques.

It can be seen from Fig. 8 that, at the lowest average bit rate without degradation of detectability by the computer, the MSE from the different methods varied from about 36 to 274. However, the MSE did not correlate with the detectability of the microcalcifications by the computer. For example, at a higher MSE of 274, the DCT-FFEC method with three- and nine-bit splitting and entropy coding of six to two bits provided a higher detectability than the LPHC technique with seven-bit quantization at an MSE of 154. The relative image quality and the information content in the decompressed images therefore cannot be judged by comparison of the MSE. Experimental measurement of the detectability of signals in the decompressed images has to be performed for both human and machine observers in order to determine the loss of image information. The acceptable degree of information loss will also be dependent on the detection task.

For the purpose of this study, the detectability of microcalcifications by computer on the decompressed images was compared to that on the original images when the same preprocessing method for SNR enhancement in the CAD algorithm was used for each image compression method. For example, for the DCT-FFEC approach, a bandpass filter used in our previous studies[4,25] was used to extract the difference image. For the LPHC method, the Laplacian pyramid images were used to reconstruct the difference image. It can be seen by comparing the highest detection curves in Figs. 6 that the multiresolution Laplacian pyramid images can provide a higher detectability than the bandpass filtered images. Therefore, although the LPHC method is less efficient than the DCT-FFEC method for image compression, the Laplacian pyramid decomposition can be useful for SNR enhancement in the CAD program. This was also the motivation that we chose to evaluate the LPHC method for image compression in CAD applications.

It may be noted that, in the LPHC method, we chose to use linear quantization for compression of the Laplacian pyramid images. It is possible that some other methods can compress more efficiently these high-frequency bandpass images and improve the performance of the LPHC method. The ranges of parameters tested in the DCT-FFEC technique were also somewhat limited. We could not explore exhaustively the different image compression methods or all possible combinations of parameters for a specific method in this study. However, our investigation did indicate the utility of the DCT-FFEC technique and the importance of proper evaluation of information loss for mammographic image compression. More extensive comparison of various compression techniques is warranted in future studies.

One principal reason that digital mammograms require an extremely high resolution is due to the potential appearance of subtle microcalcifications, which may or may not be associated with breast cancer. In general, a mammogram is characterized as an image with predominantly low-frequency contents, except for microcalcifications and subtle speculations and margin characteristics of masses, which may indicate an early breast cancer. In other words, only a very small portion of the mammogram contains clinically significant image patterns. While these patterns can be greatly distorted by an image compression technique, to our knowledge no global error measurement can quantify the effects. Many conventional compression techniques, therefore, can achieve a large compression ratio and obtain a low MSE without producing obvious visual degradation. However, these image compression techniques may degrade clinically significant information such as microcalcifications. In this study, our machine observer indicated the potential loss of detectability due to improper compression methods. The results of this preliminary study emphasize the need that special attention should be paid to the evaluation of image fidelity when compression techniques are applied to radiological images with potential subtle disease patterns. Optimization of image compression techniques based on analysis of detailed image features has recently been pursued by Lo *et al.*[26]

As found in our previous study,[4] the FROCFIT program[27] could not provide good fits to our FROC curves. Our attempt of applying the alternative FROC analysis[28] to the detection data and subsequently the CLABROC program[29] to the pair of correlated AFROC curves also failed to obtain reasonably fitted curves. This problem was probably caused by the correlation of the individual FP signals detected in an image due to the clustering criterion used in the detection process. We therefore could not use a fitted FROC curve or a single index such as the area under the AFROC curve[28] for comparison of the detection performance among different conditions. Because a rigorous statistical test of the significance of the differences between pairs of FROC curves was not yet available, we applied a paired *t* test to the TP values at a given FP to estimate the statistical significance of the differences between the detection accuracy obtained from each pair of conditions. The number of TP signals detected under the first condition for an image at an SNR threshold that yielded a given mean number of FP signals was paired with the corresponding TP signals detected under the second condition for the same image at an SNR threshold that yielded a similar mean FP. The *t* test was performed for TP pairs over a range of FP values of interest. The inclusion of TP pairs over a range of FP values took advantage of the consistency of the differences between the two FROC curves over the range of interest, similar to a curve fitting approach. However, the statistical significance might be somewhat overestimated because of the potential correlation between the TP pairs at the different SNR thresholds. An alternative test may be a paired *t* test of the differences in the partial areas under the FROC curves over the range of FP values of interest for the corresponding image pairs. The validity of these tests may be evaluated when a rigorous statistical significance test for FROC curves is developed.

V. CONCLUSION

We evaluated two image compression methods in this study. It was found that the DCT-FFEC method with bit splitting is more efficient than the LPHC method with linear quantization for compression of mammographic images, without degradation of the detectability of subtle microcalcifications by our automated detection algorithm. The highest compression ratio achieved without significant loss in detection accuracy was 9.6:1. It was demonstrated that the MSE was a poor indicator for comparison of information loss due to image compression. The evaluation of the acceptability of an image compression technique should therefore be based on human and machine observer studies.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant No. CA 48129 and U.S. Army Grant No. DAMD 17-93-J-3007 (through Subgrant No. GU RX 4300-803UM from Georgetown University). The content of this publication does not necessarily reflect the position of Georgetown University or the government, and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Diane Williams for secretarial assistance.

APPENDIX A: LAPLACIAN PYRAMID HIERACHICAL CODING (LPHC)

A schematic of the LPHC technique for image compression and decompression is shown in Fig. 9. An input image G_0 is low-pass filtered with a local and symmetric weight kernel $w(m,n)$ and then subsampled by every other pixel to different levels sequentially according to the following relationship:

$$\mathrm{REDUCE}(G_{k-1}) = G_k(i,j) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} w(m,n) G_{k-1}(2i+m,2j+n), \tag{A1}$$

where $1 \leqslant k \leqslant N$ is an index of compression level, N is the number of levels in the pyramid, and (i,j) is the pixel location in the image. The matrix size of an image at the kth level of the pyramid, G_k, is reduced by a factor of 4 compared with the $(k-1)$th level image, and is referred to as a "reduced" version of G_{k-1}. The reduced image is then expanded with a similar operation:

$$\mathrm{EXPAND}(G_k) = E_{k-1}(i,j) = 4 \sum_{m=-2}^{2} \sum_{n=-2}^{2} w(m,n) G_k\left(\frac{i-m}{2}, \frac{j-n}{2}\right), \tag{A2}$$

where the summation is performed over the terms for which $(i-m)/2$ and $(j-n)/2$ are integers. An error image in level $k-1$ is given by the difference between G_{k-1} and E_{k-1}:

$$L_{k-1} = G_{k-1} - \mathrm{EXPAND}(G_k). \tag{A3}$$

By performing the reduction N times, as shown in Fig. 9, a sequence of N low-pass filtered images with successively reduced spatial resolution and reduced sampling rate is obtained. Since one of the important convolution weight kernels resembles the Gaussian probability distribution, this se-

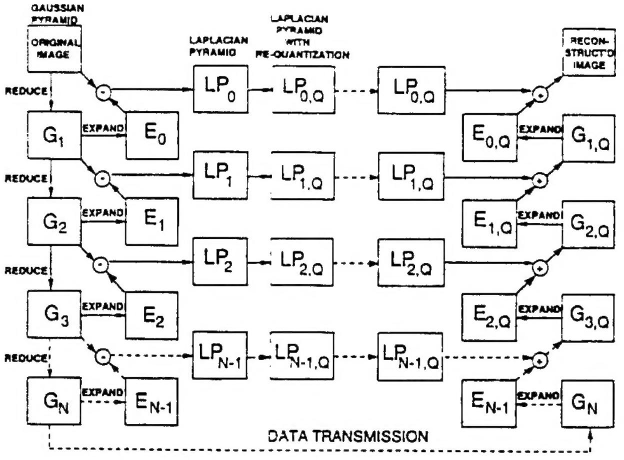


FIG. 9. Schematic diagram of the Laplacian pyramid hierarchical coding (LPHC) and decoding technique.

quence of low-pass filtered images is referred to as the Gaussian pyramid.

By expanding the reduced image and calculating the error image at each level, a sequence of N subband images of reduced sampling rate is also obtained. This sequence of error images is composed of bandpass filtered images and is referred to as the Laplacian pyramid. The scale of the Laplacian operator doubles from level to level of the pyramid, while the center frequency of the passband is reduced by an octave.

The original image can be reconstructed using the highest level Gaussian pyramid image, G_N, and the sequence of Laplacian pyramid images, L_k, $k=0,\ldots,N-1$, as shown on the right-hand side of Fig. 9:

$$G_{k-1}=L_{k-1}+\text{EXPAND}(G_k). \tag{A4}$$

If no lossy compression has been applied to G_N and the Laplacian pyramid images, the original image can be recovered without loss.

In this study, we decomposed the top level Gaussian pyramid image by the differential pulse code modulation (DPCM) technique.[30,31] The number of bits required for encoding the DPCM decomposed image was then determined by the entropy of its pixel value distribution, i.e., the histogram of its gray levels. For the Laplacian pyramid images, because their pixels values were decorrelated and could be assumed to be statistically independent, then the minimum number of bits per pixel required to exactly encode the image was given by its entropy. This optimum might be approached in practice through techniques such as variable-length encoding. The entropy of the pixel value distribution of an image was given by

$$\text{Entropy}=-\sum_{i=0}^{4095} f(i)\log_2 f(i), \tag{A5}$$

where $f(i)$ was the observed probability of occurrence of gray level i. Assuming that the variable-length code words were used in data transmission to take advantage of the non-uniform distribution of pixel values, the effective number of bits for a given Laplacian pyramid level was its entropy times its matrix size. The effective number of bits per pixel for the encoded image was thus the sum of the number of bits for all levels of the component images divided by the matrix size of the original image.

Following the approach described by Burt and Adelson,[21] a 5×5 kernel of weights $w(m,n)$ that was separable to $w(m,n)=h(m)h(n)$ was used in this study. Here h was a symmetric function, such that $h(i)=h(-i)$, for $i=0,1,2$. The weights were subject to the constraint that all nodes at a given level contributed the same total weight to nodes at the next higher level. Therefore, $h(0)=a$, $h(-1)=h(1)=\frac{1}{4}$, $h(-2)=h(2)=\frac{1}{4}-a/2$. The constant a was chosen to be 0.4 in this study to obtain a Gaussian-like function. A uniform quantization by eliminating the least significant bits was applied to the Laplacian pyramid images. Although these parameters might not be optimal choices for mammographic images, they were selected as typical values for study of the effects of the LPHC method on mammograms.

APPENDIX B: DISCRETE COSINE TRANSFORM-FULL FRAME ENTROPY CODING (DCTF-FEC)

A schematic diagram of the FFEC technique with splitting and remapping is shown in Fig. 10. An input image of $(n+k)$ bits is split into n most significant bits (MSB) and k least significant bits (LSB). Because the MSB in medical images are highly correlated, they can be encoded by correlation encoding such as Lempel-Ziv (LZ) coding or run-length/Huffman coding with high compression ratio. For the LSB image, the bits are remapped in order to convert the residual data into an image with a more continuous tone. The remapping of the LSB, denoted as RLSB, for an image with gray levels $g(x,y)$ can be expressed as

$$\text{RLSB}_k(g(x,y))=\text{LSB}_k(g(x,y)), \quad \text{for } [g(x,y)\&2^k]=0, \tag{B1}$$

$$\text{RLSB}_k(g(x,y))=2^k-1-\text{LSB}_k(g(x,y)),$$
$$\text{for } [g(x,y)\&2^k)\neq 0, \tag{B2}$$

where "&" is the "AND" operation for the bit map of the integers. The splitted and remapped image, RLSB $(g(x,y))$

TABLE I. The parameters used in the discrete cosine transform-full frame entropy coding technique.

Zone	Frequency region	No. of bits for coding	
1	0–63	Floating point	
2	64–127	12	
		Maximum	Minimum
3	128–1023	8	2
		7	2
		6	2
		5	2

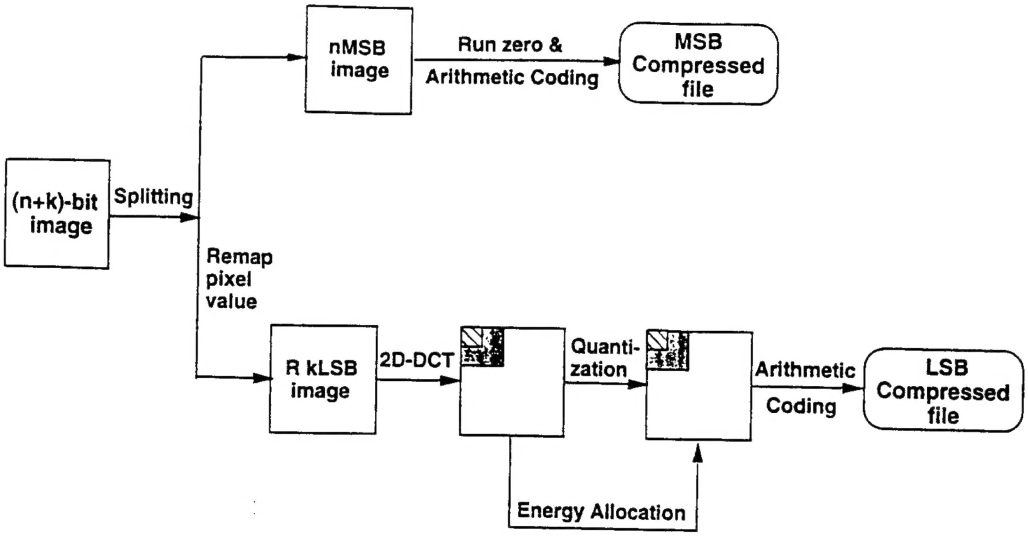


FIG. 10. Schematic diagram of the discrete cosine transform-full frame entropy coding (DCT-FFEC) technique with image splitting and remapping.

is then subject to a two-dimensional DCT. The spatial frequency domain image is divided into three zones for linear quantization. The zone boundaries and the number of bits used in each zone are tabulated in Table I. In the low-frequency zone, the DCT coefficients are stored and transmitted as the original floating point values so that no information is lost. In the mid-frequency zone, the coefficients are quantized to 12 bit integers. In the high-frequency zone, the coefficients are quantized to a range of bits specified by an input maximum and minimum number. The compression ratio is large when the maximum number of bits allowed is small. A specific number of bits to be used for a given coefficient in this zone is determined by an energy allocation scheme.[6,7] The ranges of seven to two bits, six to two bits, and five to two bits were compared for the compression schemes with splitting in this study. The quantized coefficients were then submitted to a statistical coding routine for data packing. The standard statistical coding schemes included arithmetic and Huffman coding, of which the former was used in this study.

For FFEC without bit splitting and remapping, the procedure is similar to those described above. The only difference is that there is no MSB image to be encoded. The entire image undergoes DCT, zonal quantization, and arithmetic coding, as illustrated in the lower path of Fig. 10. For energy allocation, the ranges of eight to two bits to six to two bits were evaluated.

To decompress the FFEC image, the reverse operation of compression is employed. The quantized coefficients are decoded, followed by reverse quantization, and then by the inverse DCT. Because of the quantization process that reduces the DCT coefficients of real numbers to integers with a finite number of bits, the reverse quantization cannot recover the original image information in its entirety. The degree of information loss with the FFEC technique depends on the information content of the input images and the compression ratio.

[a)] Correspondence and reprint address: Heang-Ping Chan, Ph.D., University of Michigan, Department of Radiology, Taubman Center 2910, 1500 E. Medical Center Drive, Ann Arbor, Michigan 48109-0326. Telephone: (313) 936-4357; Fax: (313) 936-9723; Electronic mail: chanhp@umich.edu
[b)] Department of Radiology, Georgetown University.
[c)] Current address: Radiological Sciences and Technology, Massachusetts General Hospital.
[d)] Current address: Department of Diagnostic Radiology, Stanford University.
[e)] Department of Radiation Oncology, University of Michigan.

[1] H. Seidman, S. K. Gelb, E. Silverberg, N. LaVerda, and J. A. Lubera, "Survival experience in the Breast Cancer Detection Demonstration Project," CA Cancer J. Clin. **37**, 258–290 (1987).
[2] D. Winfield, M. Silbiger, G. S. Brown, L. Clarke, S. Dwyer, M. Yaffe, and F. Shtern, "Technology transfer in digital mammography: Report of the Joint National Cancer Institute–National Aeronautics and Space Administration Workshop of May 19–20, 1993," Invest. Radiol. **29**, 507–515 (1994).
[3] H. P. Chan, C. J. Vyborny, H. MacMahon, C. E. Metz, K. Doi, and E. A. Sickles, "Digital mammography: ROC studies of the effects of pixel size and unsharp-mask filtering on the detection of subtle microcalcifications," Invest. Radiol. **22**, 581–589 (1987).
[4] H. P. Chan, L. T. Niklason, D. M. Ikeda, K. L. Lam, and D. D. Adler, "Digitization requirements in mammography: Effects on computer-aided detection of microcalcifications," Med. Phys. **21**, 1203–1211 (1994).
[5] S.-C. B. Lo and H. K. Huang, "Compression of radiological images with 512, 1024, and 2048 matrices," Radiology **161**, 519–525 (1986).
[6] S.-C. B. Lo, B. Krasner, S. K. Mun, and S. C. Horii, "Full-frame entropy coding for radiological image compression," Proc. SPIE **1444**, 265–277 (1991).
[7] S.-C. B. Lo, E. L. Shen, S. K. Mun, and J. Chen, "A method for splitting digital value in radiological image compression," Med. Phys. **18**, 939–946 (1991).
[8] T. Ishigaki, S. Sakuma, M. Ikeda, Y. Itoh, M. Suzuki, and S. Iwai, "Clinical evaluation of irreversible image compression: Analysis of chest imaging with computed radiography," Radiology **175**, 739–743 (1990).
[9] H. MacMahon, K. Doi, S. Sanada, S. M. Montner, M. L. Giger, C. E. Metz, N. Nakamori, F.-F. Yin, X.-W. Xu, H. Yonekawa, and H. Takeuchi, "Data compression: Effect on diagnostic accuracy in digital chest radiography," Radiology **178**, 175–179 (1991).
[10] D. R. Aberle, F. Gleeson, J. W. Sayre, K. Brown, P. Batra, D. A. Young, B. K. Stewart, B. K. T. Ho, and H. K. Huang, "The effect of irreversible image compression on diagnostic accuracy in thoracic imaging," Invest. Radiol. **28**, 398–403 (1993).

[11]J. W. Sayre, B. K. T. Ho, M. I. Boechat, T. R. Hall, and H. K. Huang, "Subperiosteal resorption: Effect of full-frame image compression of hand radiographs on diagnostic accuracy," Radiology **185**, 599–603 (1992).

[12]S.-C. B. Lo, H. Li, B. H. Krasner, M. T. Freedman, and S. K. Mun, "Full-frame compression of discrete wavelet and cosine transforms," Proc. SPIE **2431**, 195–202 (1995).

[13]P. W. Jones, S. Daly, R. S. Gaborski, and M. Rabbani, "Comparative study of wavelet and DCT decomposition with equivalent quantization and encoding strategies for medical images," Proc. SPIE **2431**, 571–582 (1995).

[14]M. Smith and S. Eddins, "Analysis/synthesis techniques for subband image coding," IEEE Trans. Acoust. Speech, Signal Proc. **38**, 1446–1456 (1991).

[15]J. Wang and H. K. Huang, "Three-dimensional medical image compression using a wavelet transform with parallel computing," Proc. SPIE **2431**, 162–172 (1995).

[16]M. A. Goldberg, M. Pivovarov, W. W. Mayo-Smith, M. P. Bhalla, J. G. Blickman, R. T. Bramson, G. W. L. Boland, H. J. Llewellyn, and E. Halpern, "Application of wavelet compression to digitized radiographs," Am. J. Roentgenology. **163**, 463–468 (1994).

[17]H. Benoit-Cattin, O. Baudin, A. Baskurt, and R. Goutte, "Coding mammograms using wavelet transform," Proc. SPIE **2164**, 282–290 (1994).

[18]H. S. Huang, L. Guan, and H. Kung, "Compression of digital mammogram databases using a near-lossless scheme," Proc. ICIP **2**, 21–24 (1995).

[19]A. J. Maeder, "Mammogram compression using adaptive prediction," Proc. SPIE **2431**, 216–231 (1995).

[20]H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).

[21]P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," IEEE Trans. Commun. **COM-31**, 337–345 (1983).

[22]H. P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. MacMahon, and P. M. Jokich, "Image feature analysis and computer-aided diagnosis in digital radiography. 1. Automated detection of microcalcifications in mammography," Med. Phys. **14**, 538–548 (1987).

[23]H. P. Chan, K. Doi, C. J. Vyborny, K. L. Lam, and R. A. Schmidt, "Computer-aided detection of microcalcifications in mammograms: Methodology and preliminary clinical study," Invest. Radiol. **23**, 664–671 (1988).

[24]P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free response approach to the measurement and characterization of radiographic observer performance," Proc. SPIE **127**, 124–135 (1977).

[25]H. P. Chan, S.-C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," Med. Phys. **22**, 1555–1567 (1995).

[26]S.-C. B. Lo, H. Li, J. Wang, M. T. Freedman, and S. K. Mun, "On optimization of orthonormal wavelet decomposition: Implication of data accuracy, feature preservation, and compression effects," Proc. SPIE **2707**, 201–214 (1996).

[27]D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," Med. Phys. **16**, 561–568 (1989).

[28]D. P. Chakraborty and L. H. L. Winter, "Free-response methodology, alternate analysis and a new observer-performance experiment," Radiology **174**, 873–881 (1990).

[29]C. E. Metz, P. L. Wang, and H. B. Kronman, "A new approach for testing the significance for differences between ROC curves measured from correlated data," in: *Information Processing in Medical Imaging*, edited by F. Deconinck (Martinus Nijhoff, The Hague, 1984).

[30]C. C. Cutler, "Differential quantization of communication signals," Patent 2-605-361, Application June 1950, Issuance July 1952.

[31]R. P. Abbott, "A differential pulse-code-modulation coder for videotelephone using four bits per sample," IEEE Trans. Commun. Tech. **COM-19**, 907–913 (1971).